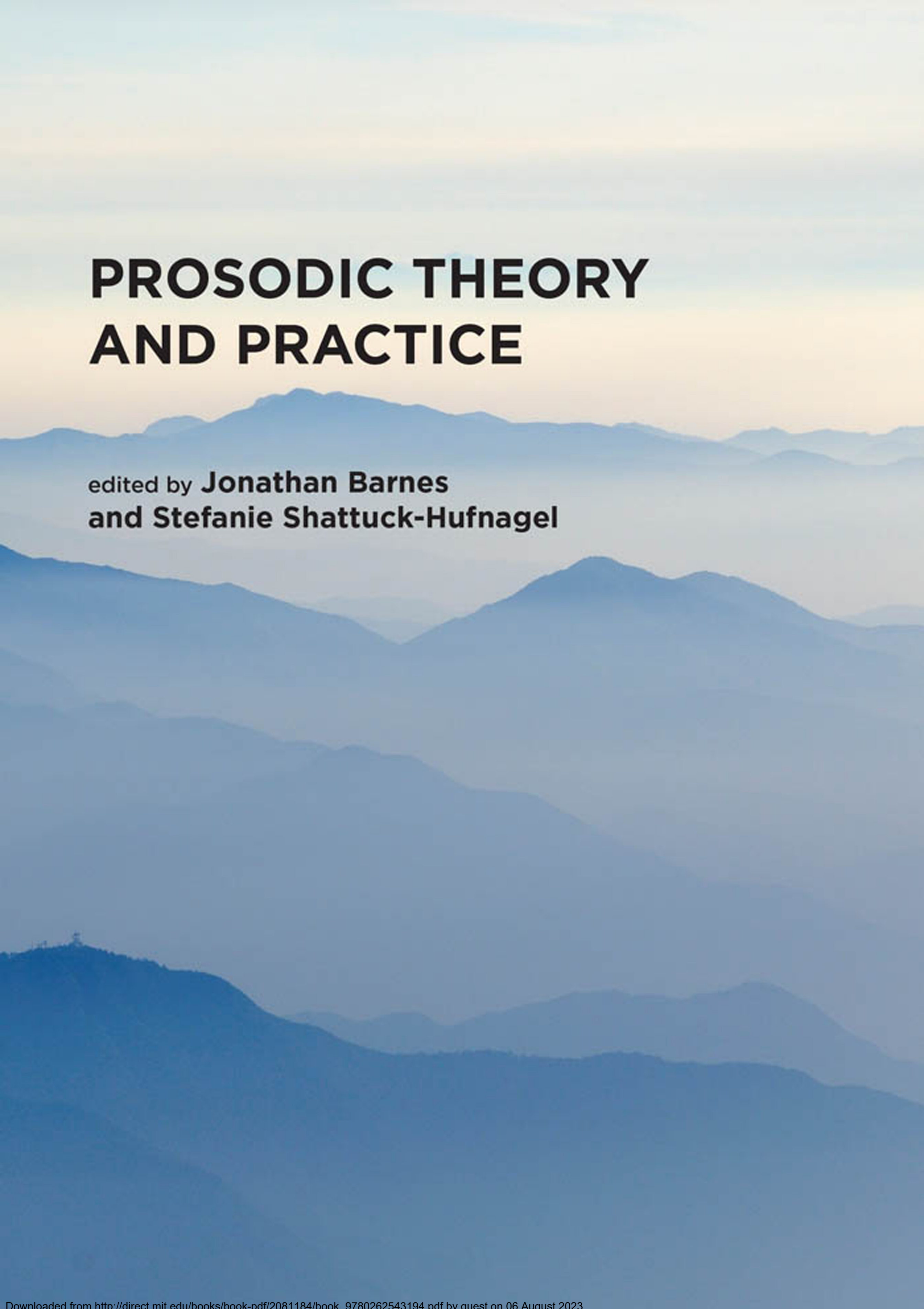


PROSODIC THEORY AND PRACTICE

edited by **Jonathan Barnes**
and Stefanie Shattuck-Hufnagel

PROSODIC THEORY AND PRACTICE

PROSODIC THEORY AND PRACTICE

Edited by Jonathan Barnes and Stefanie Shattuck-Hufnagel

The MIT Press
Cambridge, Massachusetts
London, England

© 2022 The Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data is available.
Names: Barnes, Jonathan, 1970– editor. | Shattuck-Hufnagel, Stefanie, editor.
Title: Prosodic theory and practice / edited by Jonathan Barnes and Stefanie Shattuck-Hufnagel.
Description: Cambridge, Massachusetts : The MIT Press, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2021000764 | ISBN 9780262543170 (paperback)
Subjects: LCSH: Prosodic analysis (Linguistics)
Classification: LCC P224 .P739 2022 | DDC 414/.6—dc23
LC record available at https://lccn.loc.gov/2021000764

Contents

Preface and Acknowledgments vii

Introduction: What Are Theories of Prosody For? 1
Jonathan Barnes and Stefanie Shattuck-Hufnagel

1 **The Autosegmental-Metrical Model of Intonational Phonology 25**
Amalia Arvaniti

Commentary on Chapter 1: Introducing Flexibility into Autosegmental-Metrical Phonology 64
Martine Grice

Author Response to the Commentary: Prosodic Typology and the Handling of Variability in Intonation 76

2 **Modeling Danish Intonation 85**
Nina Grønnum

3 **A Multilevel, Multilingual Approach to the Annotation and Representation of Speech Prosody 117**
Daniel Hirst

4 **The ToBI Transcription System: Conventions, Strengths, and Challenges 151**
Sun-Ah Jun

Commentary on Chapter 4: An Enhanced Autosegmental-Metrical Theory (AM+) Facilitates Phonetically Transparent Prosodic Annotation 182
Laura C. Dilley and Mara Breen

Author Response to the Commentary: ToBI Is Not Designed to Be Phonetically Transparent 204

5 **Prosody in Articulatory Phonology 213**
Jelena Krivokapić

Commentary on Chapter 5: Time to Consider Nonoscillatory, Phonology-Extrinsic Timing Approaches to Prosody in Speech Motor Control 237
A. E. Turk

6 **The Trouble with ToBI 247**
D. R. Ladd

7 **The Prosogram Model for Pitch Stylization and Its Applications in Intonation Transcription 259**
Piet Mertens

8 **The Kiel Intonation Model—KIM 287**
Oliver Niebuhr

9 **The Rise and Fall of the British School of Intonation Analysis 319**
Francis Nolan

10 **The PaIntE Model of Intonation 351**
Antje Schweitzer, Bernd Möbius, Gregor Möhler, and Grzegorz Dogil

11 **The PENTA Model: Concepts, Use, and Implications 377**
Yi Xu, Santitham Prom-on, and Fang Liu

Commentary on Chapter 11: Comparing the PENTA Model to Autosegmental-Metrical Phonology 408
Janet B. Pierrehumbert

Author Response to the Commentary: Multiple Layers of Meanings Can Be Linked to Surface Prosody without Direct Mapping 425

Index 435

Preface and Acknowledgments

The ideas that came to shape this book arose for both its editors independently, around the same time, already several summers ago. For one of us, the spark came, for whatever reason, over rice porridge, during an otherwise agreeable breakfast at a professional meeting in a major Asian city to remain nameless. The previous day's discussions, featuring some of the most thoughtful and creative prosody scholars of our time, had focused in large part on a comparison of the ideas, assumptions, virtues, and liabilities of the main theoretical approaches to prosody in active use today. Such a discussion should have been tremendously illuminating, and yet it felt at least to one observer more like a series of missed opportunities. Meanwhile, halfway across the world, during yet another workshop, the other of us was coming to a similar set of conclusions: prosody research today is moving at breakneck speed. There is more interest now, and more activity, from researchers representing an unprecedentedly broad range of disciplines, than perhaps has ever graced a subfield of the science of speech and language. Interest comes now not just from linguistics, but also from cognitive psychology and neuroscience, from computer science, speech technology (automatic recognition and synthesis of speech), and even from the medical profession, where it is increasingly recognized as a clinically significant domain.

At the same time, gaining competence in this area can be tricky. Models of often great technical sophistication have proliferated faster than even most established researchers can claim to control comprehensively. For beginning students, or scholars from allied disciplines taking first steps in the prosody literature, the situation is more dire still. While excellent surveys of the basic questions in the field do exist, they are typically couched within, or primarily devoted to, a single theoretical framework. Furthermore, as our own recent experiences in a number of public discussions at professional meetings have driven home to us, conflicting theories are often not mutually well understood. (Indeed, the extent to which competing theories even conflict in the first place is often both far from clear, and less actively investigated than one might initially suppose.) Part of the problem seems to be that central issues, such as the nature and role of phonology within prosodic theory, are understood differently by different people, in ways that are seldom made fully explicit. Criteria for the evaluation of a model's success or failure are often similarly assumed, rather than stated, leaving talented and well-intentioned proponents of competing theories sometimes talking at cross-purposes.

What would be ideal, we felt, would be if someone were to compile, in a single published resource, compact and accessible presentations of each major approach currently influential in the realm of prosodic theory and practice (e.g., in the formation of transcription systems, corpus development, etc.). Each chapter would lay out, in its developers' own words, that theory's central goals and assumptions, its strengths (what it does well), and also its weaknesses (what it is not able, or indeed not designed, to do). These chapters then could serve both as works of reference for established scholars and, perhaps more importantly, as tutorials for students just entering the field.

The present work was conceived with an aim of doing precisely this. In inviting contributions, we tried to represent both the most influential frameworks currently under active development, as well as certain historically significant frameworks. As the astute reader will surmise, we were not able to secure contributions representing all the theories we had hoped to include, but we believe the coverage we did achieve does justice to the diverse landscape of prosodic inquiry today.

As we describe in more detail in our introduction, each invited contributor was provided not just with a general topic and target word count, but also with a list of questions that we feel are critically in need of answers if we are to understand adequately how different research traditions relate to one another, and what their comparative strengths and weaknesses might be. Furthermore, to the extent that certain approaches represent subjects of particular ongoing controversy in the field at this moment, in a number of cases we have invited, as a complement to the main chapter, a critical commentary on the chapter written by a researcher whose views we felt might represent a particularly illuminating counterpoint to those of the frameworks' developers.

For the selection of commentators, one could imagine an editorial strategy that would seek to maximize disagreement between contributors. Indeed, this might seem in keeping with our current era of public spectacle vanquishing content, in which the loudest dispute is automatically assumed to be most entertaining, and hence also worthiest of our attention. This is not what we chose to do. Instead, in each case, we asked ourselves whose views on a given framework we would personally be most interested in. In some cases, this led us to avoid soliciting commentaries from researchers whose views on a given framework are widely known, or have already been published elsewhere. We think the result, together with, in some cases, authors' responses to their respective commentaries, is a series of exchanges that are both cordial and enlightening. They also represent the conversations we would most like to see continued as the field develops. To readers involved in one way or another with the production of this volume, it is likely clear that this chapter-commentary-response format has not resulted in the rapidest publication schedule one might have envisioned for this book. We hope most will agree, however, that it was worth the wait.

Long ago, before we even drafted a proposal for this book, we agreed that our primary reason for putting it together was just that we both really wanted to read it when it was done. We feel a great deal of satisfaction now that we are finally getting to do so, and can share that enjoyment with the broader community of prosodists as well. We are also extremely grateful to the many people that helped to make this book possible. Our deepest gratitude goes to the chapter authors and commentators, without whom this book clearly would not exist. Their patience and cooperation through the lengthy gestation process made the task of collecting all contributions a more reasonable one. Our colleagues at The MIT Press have been an important part of the process, and Bev Miller's editorial assistance in preparing a stylistically unified manuscript draft also provided a critical boost to our progress. In addition, a number of colleagues, including some of the chapter authors, provided additional advice and encouragement; we are especially grateful to Janet Pierrehumbert, Mariapaola D'Imperio, and Bob Ladd in this regard. We also thank the participants in a special session at a relatively recent Speech Prosody meeting, whose strikingly different views on what constitutes a theory of prosody galvanized us to create this book. Finally, we wish to express our thanks to certain people who are not represented here but whom we nonetheless regard as some of the mothers and fathers of modern prosody theory, without whom none of the work described in this volume would have been possible, including Mary Beckman, Bruce Hayes, Lisa Selkirk, Marina Nespor, Irene Vogel, and Carlos Gussenhoven. Despite the formidable challenges of understanding how speech prosody works, and of formalizing and quantifying that understanding, we are inspired by multiple discussions with a wide range of prosodists over the years to believe that this knotty set of problems will yield to theoretical insight eventually. And we hope that this volume will contribute to that process.

Jonathan Barnes and Stefanie Shattuck-Hufnagel
September 2019, Greater Boston, MA

Introduction: What Are Theories of Prosody For?

Jonathan Barnes and Stefanie Shattuck-Hufnagel

It is one thing to criticize a scientific model on the grounds that the account it provides of some phenomenon seems incorrect or incomplete. It is another thing entirely to criticize a model for never intending to offer an account of something in the first place. Often enough, reasonable scholars just disagree as to the best approach to some phenomenon, and sometimes, at least, this disagreement leads to debate that is productive and progress-oriented. Sadly, this is not always the outcome, and it is our suspicion, furthermore, that at least much of the time in prosody research, the fault for this lies primarily in a misunderstanding of the nature of the criticism being offered. Researchers, we have noticed, often use shared terminology in subtly different ways. These differences, frequently stemming from underlying differences in the goals researchers have set for themselves and their models, are also often either unacknowledged or unexplored by researchers, whose subsequent disagreements are animated precisely by these unspoken divergences.

This volume has two complementary goals, which together we hope can help to address this problem. First, we hope to provide a comprehensive overview of each of the major theoretical approaches to spoken language prosody that are currently practiced, in some cases along with critical commentaries, and responses from the original authors. Second, to deal with the aforementioned "failures to communicate" we fear we witness all too frequently in the world of speech and language research, we felt it was important to ensure that all contributions to the volume address a common set of questions that we believe are critical to a comparative evaluation of what each approach surveyed does best, and what it leaves aside as not central to its enterprise. To this end, we asked that all contributing authors give consideration to a small set of issues that we feel are not currently addressed explicitly, or in the same terms, by all approaches, and yet are generally agreed to be of central importance to the field at large. Our ultimate goal was to elicit the clearest possible treatment in each chapter of two basic, overarching questions: (i) What does the model take to be the central goal(s) of a prosodic theory? and (ii) What are the model's central assumptions (and what would it take to falsify them)? To this end, the specific issues we asked each author to address in one way or another explicitly were these:

1. *Phonology*: What is the role of phonological representations in the model? Does the model see a role for a phonological level of representation in prosodic theory, and if so, what does it look like? Does it posit a set of abstract, symbolic primitives that combine to generate well-formed intonation contours linked both to function/meaning and to the range of acceptable phonetic instantiations?
2. *Meaning*: In what way does the model connect prosodic forms to meanings? For example, is a given intonation contour (taken holistically) linked directly to a given meaning

or function? Or are the meanings of contours derived compositionally from structures built out of meaningful subconstituents? Another way of putting this might be to ask where, if anywhere, is the notion of the morpheme to be located in prosodic structure?

3. *Phonetics*: What is the relation of the model to phonetic implementation? Does the model come with an explicit theory of phonetic implementation, possibly one that could serve as the basis for synthesis of prosody? What role, if any, do the theory's proponents take synthesis to have in the evaluation of competing models of the grammar of prosody?
4. *Typology*: What is the relation of the model to prosodic typology? Does the model make predictions about the kinds of prosodic systems that should, or should not, be found in the languages of the world? What are these predictions?
5. *Psychological status*: What is the psychological status of the model? Is the model imbued with some form of psychological reality? Does it aspire to model cognition, or does it focus instead on successfully modeling or deriving the F0 contours themselves? (Or does it do both?)
6. *Transcription*: What is the relation of the model to systems of prosodic transcription? Is the model useful as a tool for prosodic transcription? If so, is it only a transcription system? If not, is it entirely without implications for transcription?

It is our hope that such an overview of each existing model, addressing at the same time a fixed set of questions identified by the editors, will provide both a useful introduction, and a handy reference guide, for a diverse audience of readers. The fact that these chapters are composed in a parallel fashion, and devoted directly to the same set of questions, will allow readers to compare and contrast each theoretical approach explicitly with all its competitors, in a way that has not to date been convenient. We also hope that the critical commentaries, provided for chapters where we detect in the literature particular levels of ongoing controversy, together with authors' responses, will provide readers with a flavor of current debate in the field, as well as an introduction to the issues most under contention today. Lastly, we hope that this direct comparison of theories on so broad a range of issues as phonological representation, phonetic implementation, prosodic typology, transcription, and corpus management, and so on, will throw into the sharpest possible relief just what is currently known about these matters, and more importantly, what remains to be discovered.

In the following sections, we will review the six questions and their motivations, with the goals of explaining why we felt that specifically these issues were the ones most in need of explicit treatment by the proponents of competing theories of prosody, and also to make maximally clear what we ourselves take the terms invoked to mean. Our aim in so doing is not to present our own views of what a theory of prosody should look like but, rather, to sketch out the range of possibilities for what one might look like, and what particular choices might mean for a given model's most fruitful application. Throughout, we give examples from particular theories, including those presented in this volume, where they most serve to clarify. But in general, we do not attempt to summarize the content of those chapters here, allowing each chapter to speak for itself. Instead, to exemplify what candidate answers to our questions might look like in the context of a single, concrete model, we focus here on one important model that is not presented independently in this volume: the Fujisaki model (Fujisaki and Hirose 1984, and subsequent publications from Fujisaki and collaborators). We have chosen the Fujisaki model for this purpose not because we take any particular stance on the correctness of its approach to the issues, but rather because, correct or not, this model's developers and practitioners have, over the years, been exceedingly and laudably clear as to what the model's aims and assumptions are, leaving comparatively little to the potential user's imagination as to where it stands on the issues we wish to address. We therefore begin this review with question 1, concerning the nature and role of phonological representations in each model.

I.1 Question 1: Phonology

Ladd ([1996] 2008, especially chapter 1) argues at length that there is a fundamental division to be made between models of intonation that are phonological in nature, and models of intonation that are not. A *phonological model*, to borrow from Pierrehumbert (see commentary in chapter 11, this volume), is one that posits a relatively small set of abstract sound categories that serve as a narrow bridge between the "extremely rich, high-dimensional world of meanings and communicative functions" and the similarly rich and nuanced universe of the acoustic signal. Such a system of sound categories would be effective in managing the flow of information between these two domains both in language learning and in daily use, in that it would allow "articulatory and perceptual patterns exhibited in one word to be reused in other words."[1] In such a system, the encoding and expression of intonation would look more or less exactly like that of segmental phonology, in the sense that the link between a given intonational category and the various meanings it plays a role in expressing would be arbitrary. That is, there should be nothing any more "declarative" about the Autosegmental-Metrical (AM) string H* L-L% in English than there is feline about the string [kʰæt] for *cat*. Furthermore, each instance of a given sound category should be subject to the same set of conditions or procedures for context-specific realization as every other instance of that category, regardless of the meaning or function it serves in a particular construction. All the allophonic variability, patterns of reduction, and so on, that might affect the production or perception of a word-initial /p/, other contextual factors held equal, should apply equivalently, whether the lexical item in question is "potato" or "Parisian." (On this view, what has sometimes been called word-specific phonetics might be usefully viewed as a form of variation dependent on detailed contextual information that includes, for example, predictability.) A *non-phonological model*, by contrast, would be one in which no such inventory of categories existed. Instead, meanings or functions would map in one way or another directly to the acoustics of the signal.

The Fujisaki model is perhaps best thought of as a model of the mapping between phonological representations, about the precise nature of which it is largely agnostic, and the acoustic signal. Fujisaki and Hirose (1984) characterize it as a model of the "control process" of the fundamental frequency. It is a quantitative model, aspiring to accurate, nuanced description of the F0 contour, but crucially, in a manner that reflects the relationship of the control mechanism to the discrete linguistic categories that drive it. While the model took as its original impetus a particular linguistic analysis of Japanese prosody, it has since been extended to model F0 patterns in other languages as well, and, in principle, would be compatible with a variety of phonological analyses of the languages to which it has been applied. Mixdorff and Fujisaki (2000), for example, explore the use of Fujisaki model parameters as a means of mapping between the acoustic signal and a symbolic representation of the kind embodied for German by the AM **To**nes and **B**reak **I**ndices (G-ToBI) system.

The Fujisaki model may not be directly committed to a particular model of phonology, but it does have implications for the structure of any such model. These implications have at times drawn criticism from practitioners of the AM approach (e.g., Ladd [1996] 2008), and while there certainly are differences between the assumptions of the Fujisaki model about linguistic prosody and the structures normally assumed by AM theorists, a closer comparison also reveals greater commonality than is usually assumed. To see this, though, it is first necessary to explore in a bit more detail the structure of the Fujisaki model.

The bedrock structural principle underlying the Fujisaki model is that the F0 control mechanism involves the superposition of information flowing from two distinct channels. The first of these represents the global shape of the F0 contour over a relatively long, linguistically demarcatable domain, and is called a *phrase command*, while the second represents local pitch events, realized on a shorter timescale, which are called *accent commands*. Events in both channels modulate a "base frequency" taken as input, and the superposition of their effects on that base is meant to yield the specific shape of any given F0 contour. The much-reproduced diagram in figure I.1 depicts the basic structure of the model.

In the original analysis of Japanese, the accent commands were meant to implement the lexical pitch accents of that language, while the phrase commands were meant to represent downtrend (in practice, conflating what might otherwise be distinguished as global declination and local downstep). Because of the nature of Japanese prosody, the original model required neither globally upward-moving phrase components nor multiple accent commands of contrasting types or compositions. In practice, however, nothing would prevent the model from adopting such innovations, as indeed it has, in application to other languages. To model lexical tone in Mandarin, for example, as shown in figure I.2, syllables can be associated with a positive tone or accent command (tone 1, high), a sequence of negative and positive commands (tone 2, rise), a single negative command (tone 3, low), or a sequence of positive and negative commands (tone 4, fall) (Fujisaki et al. 2005). Similar enrichments can yield the contrasting pitch-accent types familiar from the tone and intonation systems of various European languages to which the model has also been applied.

It is tempting to see in the negative and positive accent or tone commands of the Fujisaki model direct analogues of the phonological Lows and Highs that make up representations in the autosegmental tradition, and indeed, in certain analyses, such as the Mandarin one in figure I.2, some parallelism does emerge (H versus LH versus L versus HL). In other cases, however, such as the standard autosegmental analysis of certain English or German pitch accents, the Fujisaki model posits only a positive accent

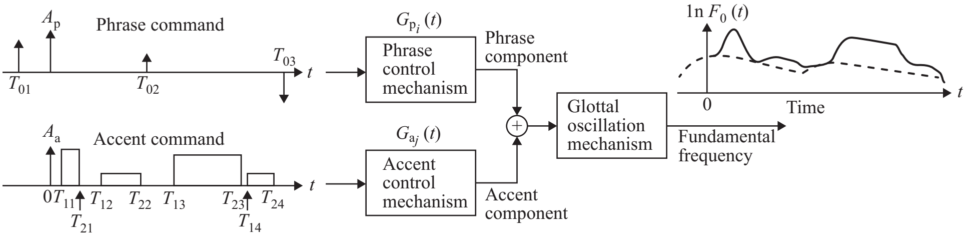


Figure I.1
A functional model for generating F0 contours of sentences. *Source*: Courtesy of Fujisaki (1997).

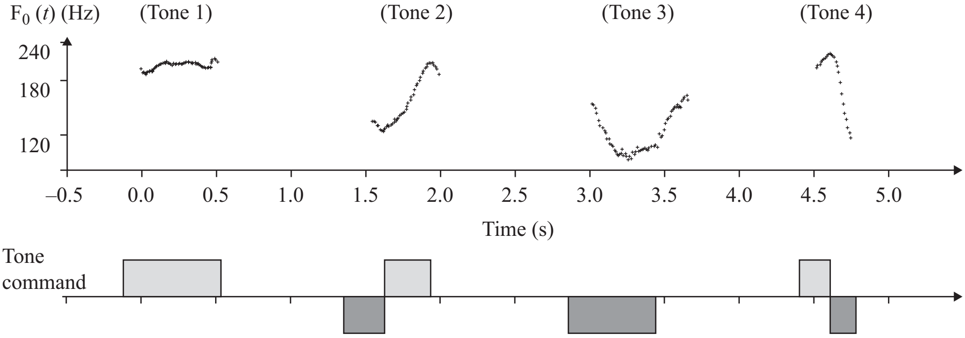


Figure I.2
Fundamental frequency contours for the four Mandarin lexical tone patterns (top) and the underlying tone command patterns used to model them (bottom). *Source*: Courtesy of Fujisaki et al. (2005).

command, where AM models need both a Low and a High to capture the sharp local rise associated with the accent (Fujisaki, Ohno, and Wang 1998). Where the parallelism breaks down even more obviously, though, is with the Fujisakian phrase command, which has no direct analogue in autosegmental terms. The phrase command is commonly understood, even by its proponents, to interpret a "global" component of the F0 contour. Therefore the phrase command appears to be something antithetical to the spirit of the standard AM approach, which purports to view tonal representations of all kinds, whether associated with prominences (e.g., pitch accents) or phrase boundaries (e.g., boundary tones or "phrase accents"), as equivalently just a sequence of locally instantiated tonal targets, instructions to reach a specific F0 level by a specified point in time.

But if Fujisaki's phrase commands are global in the sense that they determine the characteristics of the F0 contour over a relatively longer duration, they are not global in the sense that they directly encode a target shape for the F0 contour over a specific duration (as do the "grids" posited by Gårding 1983, or the "superordinate intonation contours" described by Grønnum in chapter 2, this volume). Rather, both the phrase and accent components of the model are implemented as strings of locally deployed "neuromotor commands." The effects of the phrase component on the F0 contour are modeled as the response of a critically damped second-order linear system to an impulse with a particular timing and amplitude. The longer timescale of these events in comparison with that of the accent commands is due to the response characteristics of the system, as it decays back to a specified "base frequency." The phrase impulse itself, however, is not extended in time. In some ways, it is Fujisaki's accent command that explicitly imposes particular F0 characteristics over a temporally extended domain, insofar as it is modeled as a step function that requires for its implementation explicit times of onset and offset. This is especially clear in figure I.3, where the phrase and accent components for the analysis of a sample utterance of Japanese are depicted separately.[2]

If it isn't globality per se, then, that makes AM and the Fujisaki model incompatible, perhaps it is some other aspect of the notion of *superposition*, or the division of the control mechanism into multiple distinct levels. But this too doesn't obviously divide the approaches. Superposition in the Fujisaki model does mean that the F0 at any particular moment in an utterance can represent the interaction of commands from

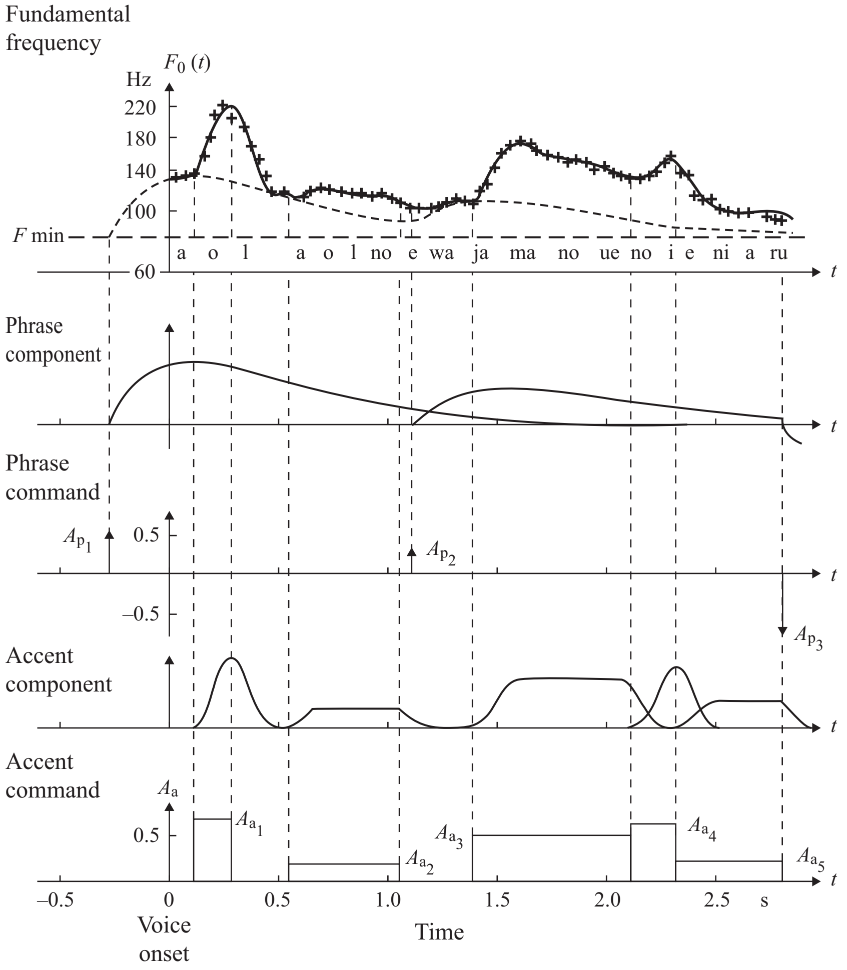


Figure I.3
Decomposition of the F0 contour for the Japanese sentence /aoi aoi no e wa jama no ue no ie ni aru/, 青い葵の絵は 山の上の家にある (The picture of a blue hollyhock is in a house on the top of the mountain) into phrase and accent components, along with the underlying commands that give rise to them. *Source*: Courtesy of Fujisaki (1997).

several distinct representational objects (rather than the apparent sequential movement from one discrete target to the next that ostensibly characterizes AM phonetic interpretation). But this is hardly an unusual situation in the phonetics-phonology interface. Indeed, even in pre-autosegmental days, when phonological representations consisted more or less exclusively of strings of symbols in strict linear arrangement with respect to one another (as per, e.g., Chomsky and Halle 1968), it seems to have been accepted as unproblematic that the state of the vocal apparatus at any given moment should simultaneously reflect the influence of multiple elements of those strings. The symbolic representations of consonant and vowel features relating to place of articulation may not overlap with one another, but measured F2 values during a given vowel, for example, will simultaneously reflect a number of factors, including "inherent" targets projected by the vowel itself, coarticulation with neighboring vowels and neighboring consonants, modulations caused by prosodic factors related to timing and prominence, and so forth.

Autosegmental representation of phonological features only amplifies this possibility, in that features share hosts (e.g., timing slots, root nodes) in ways that blur the edges of the segment as traditionally understood. Indeed, the whole notion of a tonal tier that exists in parallel to the representation of the segmental melody is nothing if not superpositional, in the sense that both the tonal targets, and the various laryngeal states demanded by the consonants and vowels of an utterance, must ultimately be realized simultaneously, by a single larynx. In some ways, then, perhaps the Fujisaki model is more autosegmental in spirit than AM itself. Perhaps the phrase and accent components could fruitfully be understood as analogous to autosegmental tiers, each housing distinct, but related kinds of pitch specifications. (C-place and V-place features come to mind [Clements 1991], as do, closer to home, distinct register and tone tiers [Yip 1980].) Ultimately, in both models distinct streams of information must eventually be conflated, yielding the temporal overlap of multiple distinct information sources that is typical of phonetic realization more generally.[3]

Regarding the need to posit two distinct types of pitch specifications in the Fujisaki model, one reason that AM models are able to achieve the appearance of representational uniformity is just that they effectively fail to provide any uniform treatment of the phenomena that ultimately motivated the phrase component of the Fujisaki model in the first place. Downtrend, reset, and various other phenomena involving baselines and register have been treated in various ways by different AM theorists, but they are more often than not simply absent from the symbolic representations of intonation contours in those models, with the understanding that they must arise somehow phonetically, through the interpretation of prosodic phrase structure, metrical relations, tonotactics, or some combination thereof. Even excluding these phenomena from consideration, though, AM representations of prosody are not always interpreted as uniformly as the target-and-interpolation approach suggests they should be. While pitch accents and certain boundary tones sometimes seem to project unique, localizable phonetic targets (identified with "turning points" in the F0 contour), the phrase accents of some analyses (Pierrehumbert 1980; Grice et al. 2000) seem to be governed by realizational principles of a different stripe, involving targets extended over broader temporal domains. While various explanations have been offered for this behavior (among them, autosegmental spreading, copying, secondary association), it still seems odd that it is this one category of tonal event that is so consistently in need of such devices. None of this is to suggest that either model is "correct" in its assumptions, but rather we are pointing out that AM models may not be so representationally uniform, nor Fujisakian superposition so alien to them, as is commonly assumed.

In sum, a phonologically based model of prosody will include a set of abstract symbolic sound categories, themselves without meaning, that can be combined into strings subject to the restrictions of the grammar, and that serve as a narrow bridge between the rich universe of meanings and the similarly rich universe of sound shapes. In contrast, a phonology-free model may map meanings directly onto sounds, without an intervening set of abstract categories, or it may simply remain agnostic as to the nature of the categories involved in phonetic implementation.

I.2 Question 2: Meaning

The second question we posed to contributors concerned the place of meaning in their prosodic models. We suggested in the formulation of the question that what we really wanted to know was where, if anywhere, the notion of the *morpheme*, or minimal meaning-bearing element, resides in their approaches. Using question 1 as a springboard here, if there is such a thing as intonational phonology, and if that thing is similar in nature to segmental phonology, then we must expect the tonal primitives of the system to be themselves devoid of meaning, just as individual consonant and vowel phonemes are said to be in non-sound-symbolic vocabulary in the classical Saussurean sense. Those phonological primitives only are associated with meanings once they are combined, according to the constraints imposed by the phonological grammar, into strings of varying lengths, which in spoken languages become the signifiers that join together in arbitrary but inviolable union with their signifieds, to make up the linguistic sign. This duality of patterning (Hockett 1958; Ladd 2014) or double articulation (Martinet 1949) has been said to be fundamental to the structure of human language in general, and is one of the more frequently mentioned properties that may (or may not) separate human language from other animal communication systems. (E.g., Cheney and Seyfarth's [1990] vervet monkeys could be argued to have mastered arbitrariness, but apparently still lack double articulation.) It is this "design feature" of language that allows the creation of the large, potentially infinite vocabularies characteristic of human languages out of small, finite sets of phonological elements (cf. Sandler et al. 2011 on the emergence of this property in linguistic systems, such as new sign languages, that may not originally have it).

Speaking only of intonation, it is perhaps not clear whether such a feature is necessary to the expression of the meanings encoded by the system, in part because it is still not clear how numerous, or even how enumerable, the meanings in question are. Still, it is somewhat surprising that in the AM tradition, at least in one of the most influential approaches, that of Pierrehumbert and Hirschberg (1990), the traditional roles occupied by the phoneme and morpheme appear to be conflated. In that approach, the individual tones making up the tone string (e.g., H*, L-, L%), potentially down to the component targets of bitonal pitch accents (e.g., the leading L of L + H* or the trailing H of L* + H [301]) are associated each with their own elements of meaning. Meanings conveyed by entire utterances are then argued to be built up compositionally from those elements, rather like sentence meanings are composed from the combinations of the morphemes they contain into particular structural configurations.[4] But if individual tonal targets are effectively morpheme-like, they are not themselves obviously composed of any phoneme-like subparts, which would appear to make intonational phonology quite different from the rest of phonology, in that it does not seem to show duality of patterning. Some of these issues are discussed interestingly by Liu et al. (2013), where it is argued that "prosodic functions," such as focus and modality, are the morpheme-like elements in intonational meaning and that their formal expression must therefore be something more akin to the whole contour than the individual tonal targets (see also Xu, Prom-on, and Liu, chapter 11, this volume, for an exposition of the model assumed in that study).

To continue giving sample answers to our questions using the Fujisaki model illustratively, here the Fujisaki model has no particular commitments that we are aware of. As a model of phonetic implementation, it assumes as its input some form of "linguistic information," which Fujisaki (1997: 28) describes as "symbolic information that is represented by a set of discrete symbols and rules for their combination." (In this instance, he has in mind the accent types of Japanese words, but one could easily substitute Mandarin tones, English pitch accents, or any other F0-based linguistic entities.) It is fair to say that capturing how these entities relate to the expression of meaning is not among the central goals of the model.

In sum, two key questions about how a prosodic model relates to meaning are whether the meaning-form mapping includes an intermediate symbolic representation in the shape of elements that themselves do not bear meaning, and whether the meaning-bearing elements are individual tonal targets, strings of targets, or some other type of unit.

I.3 Question 3: Phonetic Implementation and Synthesis

Phonetic implementation is perhaps the one area where all the models included in this volume have some explicit position or stake, and thus it might appear to be the single area most promising for direct "head-to-head" comparison or evaluation of competing models of prosody. Even in this area, though, differing aims and different assumptions make the relative "success" of the models less directly evaluable than is commonly assumed.

Once again (and if this book as a whole could be assigned a single take-home message, this would be a strong candidate), which model of prosody is seen as "correct" or "best" depends on what the user wants to do with it. No one model may answer all possible needs. For example, a model that aims to understand prosody in the context of human cognitive architecture might also yield insights about cross-language typology of prosodic systems, but then again, it might not. (The argument concerning the sources of cross-language patterns in phonology—that is, whether they are best thought of as directly encoded in grammar or linguistic cognition, or whether they might instead arise as by-products of the transmission process, from facts about articulation, speech acoustics, domain-general properties of the auditory systems, and so on—is presumably as relevant in prosody as it is in segmental phonology.) The model that currently yields the most lifelike synthesis might similarly have little to tell us about cognition, and so forth. Convergence in these domains of inquiry is possible, and even perhaps desirable, but it is not logically necessary.

The Fujisaki model is particularly rewarding to examine in this connection, because its aims in this regard are both exceptionally clearly stated in a host of publications (e.g., Fujisaki and Nagashima 1969; Fujisaki and Hirose 1984; and Fujisaki 2004, among others), and interestingly divergent from the aims of some of the models included in this volume, even where they might initially appear to be quite similar. The Fujisaki model was "conceived and developed primarily for use in (high-quality) speech synthesis" and was not originally concerned with capturing the cognitive aspects of language prosody. At the same time, it is clear that from its earliest incarnation it had the ambition not just to produce realistic-sounding synthetic speech for technological applications, but also to provide a tool by which we might learn things about the properties of human languages. Both the architecture and implementational details of the model are "deeply based on the physiology and physics of the mechanisms and processes of controlling the frequency of vibration of the vocal folds." It therefore "models a part of the process of speech production" (Hiroya Fujisaki, personal communication, 2016).

It is worth pausing at this juncture for a moment to appreciate how truly, deeply unnecessary this characteristic is, if our primary aim is nothing more than to allow electronic devices to produce synthetic utterances with natural-sounding prosody. The success of the Fujisaki model in producing high-quality synthesis shows us that it is possible for such a model to succeed, but there is no a priori reason, as far as we know, to assume that the best means for production of natural-sounding synthetic speech using a digital computer will require formulae that are conceptually parallel in any way to the control mechanisms for fundamental frequency in the human vocal tract. It is logically possible, for example, that the best synthetic prosody would in fact be produced by a model whose properties had no parallels at all to human physiology, human cognition, common grammatical properties of human languages, or anything else we might find interesting in the realm of prosody. It is entirely possible that the model that proves most useful from a technological point of view will also prove least interesting from a scientific one.

Fujisaki's goal, by contrast, is not just to synthesize natural-sounding Japanese, but in so doing, to learn something about the nature of human language. By structuring the model in such a way as to "quantitatively simulate the control process," we may then be able to use the model to "separate the characteristics of the physical and physiological mechanisms from the factors that carry linguistic information" (Fujisaki and Hirose 1982: 69). To the extent that those physical mechanisms are common to all members of the species, any differences between languages that emerge through this modeling may be assumed to give us insight into structural differences between the languages in question. Speaking more generally of what makes for a successful scientific model, Fujisaki says:

> In my humble definition, a model is a description (preferably in objective, physical, and mathematical terms) of the essential structure of a mechanism (a structural model), or the essential function of a process (a functional model). Thus it is not a subjective statement or description, but is an objective description, most preferably by mathematical terms. Note, however, that it is different from *a mere approximation by some arbitrary mathematical functions*. In my humble personal view, mathematics is a tool to describe fairly precisely what we observe in physics, physiology, or in cognitive psychology, but should be based, not on surface phenomena, but on the underlying mechanism or process that produces the observed phenomena. (personal communication; emphasis added)

In a similar vein, Hans-Jörg Mixdorff argues, concerning the insistence on "linguistically meaningful model parameters" in the Fujisaki model:

> An arbitrary number of commands provided, any F0 contour can be approximated with unlimited accuracy. For this reason constraints must be applied in order to ensure a linguistically meaningful interpretation of the analysis results. These constraints are language specific and concern the relationship between linguistic units and structures (prosodic phrases and accents for instance) and the phrase and accent commands. (1998, 51)

The phrase and accent components of the Fujisaki model are thus not adopted with their particular mathematical characteristics just because they exhibit success in modeling the F0 contours observed in natural speech, but these components are in fact identified with distinct aspects of the physiology of F0 control as well. Specifically, the activity of the phrase component is said to reflect horizontal translation of the thyroid cartilage stemming from the activity of the pars obliqua of the cricothyroid muscle, while the accent component reflects rotation around the cricothyroid joint caused by contraction of the pars recta of the same muscle (Fujisaki 1981, 1988). The differing timescales of the two components of the model, and hence the linguistic phenomena they are selected to implement (global or phrasal versus local or accentual), are thus argued to follow from the temporal properties inherent in the control of these two degrees of freedom in the movement of the thyroid cartilage relative to the cricoid. Whatever else these aspects of the model might be, then, they are clearly not arbitrary.

Fujisaki's work is also noteworthy in being among the first to invoke the notion of analysis-by-synthesis as a procedure for scientific investigation in the area of prosody. This is unsurprising, because Professor Fujisaki was present and in collaboration with researchers at MIT when the theory of analysis-by-synthesis was being developed. Indeed, he was second author on one of the earliest publications of research involving the theory (Bell et al. 1961). Since that time, outgrowths of the original analysis-by-synthesis concept have become important, and widely cited, in the prosody literature. We have noticed, however, a great deal of diversity in what different researchers seem to mean when they invoke this term today, so much so that, as with other terms (e.g., *phonology*) that have become commonplace in our discourse, we suspect that the comfort of shared terminology may be seducing us into the belief that we understand one another's ideas better than we in fact do.

With regard to the term *analysis-by-synthesis*, the original concept was first published, as far as we know, in a 1959 paper by Halle and Stevens. Described in that paper as a "dynamic system for the analysis of signals," (1) the theory was apparently intended to be used in two distinct but related ways. First, it was proposed as a model of a parsing strategy—of what "receivers" (in spoken language, listeners) may be doing when they decode linguistic signals of various kinds. Second, it was intended as a tool for use in the then quite new field of automatic speech recognition. Early in their paper, Halle and Stevens characterize the speech signal as a "particular value of some function," which they refer to as the "encoding function." The argument of this function they refer to as the "message," and therefore they represent encoding and decoding as involving the following relation (1):

$$\text{Signal} = F(\text{Message})$$

A notable property of most linguistic communication, they continue, is that, because every listener (or decoder) is at least in principle some of the time also a speaker (or encoder), listeners, when faced with the task of decoding a given signal, have access to both the same encoding function and the same basic inventory of possible messages that the speaker has. What Halle and Stevens suggest, therefore, is that the decoding process may involve a kind of "active internal replication process" on the part of the receiver. Faced with a particular spoken signal, in other words, a listener takes a set of candidate intended messages, runs those through the encoding function himself or herself, and compares the resulting internally generated signals to a stored version of the original. The candidate message yielding a signal with the least amount of error relative to the original is assumed to be what the speaker intended to communicate.

The same sort of procedure is proposed as a tool in the automatic analysis of a remarkably broad set of different kinds of linguistic signals (e.g., vowel formants, handwriting, Russian morphology). For example, Bell et al. (1961) propose a method to use analysis-by-synthesis as a step toward the automatic recognition of non-nasal vowels, by using the method to figure out the vocal tract resonances and source characteristics that were used to produce a particular spectral slice taken from the signal. As the authors of that paper explain, however, this is possible only because of the explicit models of the encoding function that we have, in the form of the acoustic theory of speech production (Fant 1960). The Signal to be decoded is the vowel spectrum sampled at a particular point in time, and the Message is "a tabulation of the resonant frequencies and relevant data concerning the excitation." (3) Further, Halle and Stevens (1959) propose that this same procedure could be used in the "second stage" of the work of a speech analyzer, this time taking those same resonant frequencies and source characteristics to be the signal, from which could be decoded a message consisting of a string of "discrete phonetic symbols." (3)

In yet another application, it is reported that Halle was at work with a colleague on a system for the automatic parsing of the morphological structure of Russian words. Halle and Stevens state explicitly that the reason such a morphological processor is even conceivable is that we believe we have a good understanding both of what the complete set of possible elements in a Message would be (the inventory of Russian morphemes), and also of the encoding function, or rules used by speakers for those morphemes' combination. It should be clear that if we lacked either of those elements, automatic morphological parsing of Russian words would not be possible using the analysis-by-synthesis method.

As with the model of vowel perception, the work of a human listener in morphological processing might proceed analogously to the proposed automated system. Because the listener and the speaker share both an inventory of possible input morphemes and the rules for their combination, a Russian listener confronted with a particular spoken lexical item might think "What set of morphemes, in what structural relationship, would it take to yield a word comparable to the one I just heard?"

In the area of prosody research, analysis-by-synthesis entered researchers' lexicons very early on. By 1965, it appears in the title of a paper by Öhman and Lindqvist ("Analysis-by-Synthesis of Prosodic Pitch Contours"), which does not, for whatever reason, cite Halle and Stevens. (Oddly, the phrase also does not appear in that paper anywhere other than in the title, making their precise conception of it difficult to ascertain in retrospect.) Fujisaki, however, who describes the genesis of his own model in large part as a response to certain aspects of Öhman's model, uses *Analysis-by-Synthesis* (which he tends to capitalize) in more or less the same way that Halle and Stevens (1959) and Bell et al. (1961) used it. Because he has an explicit quantitative model of Japanese prosody, including an encoding function, and a hypothesis about the kinds of elements that serve as its inputs, he describes how analysis-by-synthesis can be used to analyze a given Signal (a raw F0 contour) in terms of the Message (phrase commands, accent commands, and their corresponding parameter settings) that must have given rise to it. In Fujisaki (1997), he also talks about a second stage of analysis, just as Halle and Stevens did. If the first level of analysis infers the phrase and accent commands from the characteristics of the signal, then the second application of analysis-by-synthesis infers the "units and structures of prosody" (30) from the commands. Obviously, this last level of analysis is only possible assuming we have some idea what those units and structures are, and what the encoding function that maps from them to a particular set of phrase and accent commands is like.

There is, however, another common use of the term *analysis-by-synthesis* in the prosody literature, and that is as a method for the validation of scientific models. Daniel Hirst (2011), who has employed the concept of analysis-by-synthesis in his work since at least 1980, summarizes this take on the model neatly:

> The analysis by synthesis paradigm is potentially an attractive one for linguists, since it provides an empirical solution to the problem of validating an abstract model. The interaction between linguists and engineers has always been a productive area of exchange. This

> is particularly evident in the area of speech prosody. If the representation derived from a model can be used as input to a speech synthesis system, and if the contrasts represented in the model are correctly rendered in the synthetic speech, then the representation can be assumed to contain all the information necessary to express that contrast. (58)

This notion is clearly present in the original papers too, though it appears more as an aside than as an end in itself. Bell et al. (1961), for example, remark that "once a set of parameters is found such that a good replica of the input signal is generated when these parameters are applied as instructions to the internal generative model, then there is little question that this set constitutes an adequate representation of the input." (1735) Fujisaki (1997) says of his own model, again in passing, that "the close agreement of the model's output with the measured F0 contour, found in this as well as in a number of speech samples analysed, attest the validity of the model." (35)

Unfortunately, it is all too easy to take this observation a step further, and to infer that if good synthesis validates a model in some sense, then better synthesis indicates a superior model. Of course, within the confines of a single model, with a given encoding function, and a particular set of candidate input elements, analysis-by-synthesis tells us that the analysis most closely reproducing the signal is the best one. One set of phrase and accent commands, for example, and not another, might be the best decoding of a given F0 contour. But beyond this, the notion of model comparison through analysis-by-synthesis breaks down. We might in principle want to decide, for example, which theoretical model of prosody was "better," the AM model or the Fujisaki model, by comparing the quality of synthesis produced using the two. This might be a good idea, or it might not be. Either way, though, it would not constitute an application of the principle of analysis-by-synthesis. The reason should be clear, returning to that basic relation, Signal = F(Message). To analyze a given signal, the AM model and the Fujisaki model (one imagines) would employ completely different encoding functions, as well as entirely different sorts of representational primitives for their "messages."[5] By the technique of analysis-by-synthesis, given a shared encoding function, and a shared inventory of possible messages, we can determine which of two (or more) messages was more likely to have produced the signal in question. Without that common ground, though, no such inferences are possible.

We might still feel, nonetheless, that quality of synthesis is a suitable criterion by which to decide between prosodic theories, even if calling this procedure analysis-by-synthesis is historiographically misleading. Again, in this instance, we would simply invoke the caveat that in order to decide which of two models is "better," it is important first to answer the question: Better for what? Though as we understand it, this may no longer be the case, there was a moment in the history of the field of speech synthesis when the best programs for producing synthetic speech were based on concatenation of units called "diphones." *Diphones* are essentially transitions from one segment, traditionally conceived, to another. Rather than concatenating a series of consonants and vowels, in other words, diphone synthesis would concatenate a unit containing the transition from a particular consonant into a particular vowel, with the transition from that same vowel into the following consonant, and so forth. Despite the successes of diphone synthesis, however, few linguists were moved to posit that perhaps diphones were the correct representational primitives to be used in our phonological models as well. To the extent that diphones did not replace traditional consonant and vowel segments as elements of phonological representation, this was presumably because phonologists felt that whatever was gained by their adoption for purposes of synthesis was offset by the insights that would be lost concerning the phonological structures of individual languages, cross-language phonological typology, and so forth. Their aims, in other words, were different.[6]

I.4 Question 4: Typology

The purpose of this question is to determine the extent to which the prosodic models included in this volume address questions of "possible" and "impossible" prosodic systems, cross-language frequency of prosodic systems or processes, and so on. For as long as there have been phonologists, there has been a general acknowledgment of responsibility among them to questions of why certain kinds of phonological patterns (vowel inventories, rule interactions, stress systems, and so on) should be common (if not universal), while others should be rare (if not unheard of).

What constitutes an "explanation" for such cross-language patterns is the subject of much discussion in the field, the sum of which cannot be adequately reviewed in this space. (See Hyman 2018a and Gordon 2016, chapter 2, for two illuminating, though quite different, overviews of these issues.) For our purposes, it is enough to note that in some research traditions—for example, most approaches to phonology within the umbrella of Optimality Theory (Prince and Smolensky 1993)—the "typological coverage" that a proposed theoretical innovation achieves is a central, even the central, criterion by which its fitness is evaluated. Both *overgeneration* of language patterns (that is, predicting the existence of unattested grammatical systems) and *undergeneration* (failure to predict attested systems) are considered shortcomings from this point of view (though in practice, overgeneration is usually understood to be a lesser sin than undergeneration, because currently unattested language types could nonetheless turn out to be possible). If typological coverage is a common goal for the proprietors of linguistic theories, though, it is by no means a necessary one. Researchers disagree as to which, if any, cross-language regularities should be encoded in "the grammar," where that term usually, if not always, carries cognitive implications. One place this can be seen is in recent controversy over the relative roles of synchronically versus diachronically oriented explanations for typological patterns. Synchronic explanations commonly invoke what Moreton (2008) calls *analytic bias*, a cognitive predisposition, whether language-specific or domain-general, that makes a particular pattern difficult or impossible to learn or predisposes learners to favor one kind of pattern over another. Diachronic explanations, by contrast, focus on what Moreton called *channel bias*—some aspect, in other words, of language use or transmission that makes phonological changes giving rise to one kind of pattern more likely or more common than others. Synchronic explanations also tend to involve hypotheses about "higher-level" aspects of cognition, focusing on facts about learning strategies or computational complexity, while diachronic explanations tend to prefer "lower-level" or entirely noncognitive approaches, featuring instead facts about speech production, speech acoustics, audition, or perception. Thinking of synchronic explanations as strictly cognitive, and diachronic ones as noncognitive, though, is not quite right. Much of the literature on analytic bias focuses on the possibility that at least some of that bias is "substantive" in nature, meaning that humans are predisposed to learn patterns that are "phonetically natural" (where *natural* is usually taken to mean functionally preferable in light of facts about articulation, audition, and so on). In addition, as Kiparsky (2006) points out, sound change is obviously shaped by a cognitive or grammatical component as well. Opinion currently seems settled, following Moreton (2008), that both analytic and channel biases have a role to play in shaping typology. There is also some evidence (e.g., Moreton and Pater 2012; Glewwe et al. 2018) against the existence of *substantive bias*, or the preference in language learning for patterns that are phonetically natural. Many key questions, in any case, have yet to be resolved.

Our purpose in discussing this at all is just to point out that it is logically possible to wish to devise a cognitive model of how, for example, prosody works, but to believe that much or all of what we think of as prosodic typology has no place in the development or evaluation of that model. Likewise, it is entirely possible that the most successful approach to the synthesis of natural-sounding prosody will have nothing whatsoever to tell us about typology. (Obviously, in both these cases, the opposite could also be true.) What we wish to emphasize is that complaints that a psycholinguistic model fails to account for typological patterns will be constructively received only if the proponents of the model in question believe it is their responsibility to account for facts about typology in the first place. Complaining that a typologist has failed to deliver up a functioning system for prosodic synthesis may be similarly beside the point. As with the preceding three questions, though, we fear that in many instances, most of us are simply unaware of the commitments of many models in this regard, and prefer instead to hold every model tacitly responsible for whatever interests us personally the most in prosody, whether that forms part of that model's intended purpose or not.

To continue our extended case study, the Fujisaki model is not explicitly, centrally typological, but, as should be obvious from our description of its properties above, it does make clear predictions about what prosodic systems should or should not be like. The combination of phrase and accent commands with the full range of their potential parameter values will generate F0 contours of some shapes, and will not generate others.

Application to a variety of typologically distinct languages has always played a role in the model's development, and it has also at times been criticized (e.g., Ladd [1996] 2008, 28–29) for undergenerating attested F0 contour shapes. We are reasonably certain, furthermore, that if such criticisms proved correct, the models' proponents and developers would view that as a shortcoming, and seek to remedy it. More deeply, though, the very proposal that F0 events come in two distinct control varieties, corresponding to the phrase and accent commands, has ramifications for the predicted typology of human language prosody.

I.5 Question 5: Psychological Reality

At least since Sapir (1933), if not much earlier,[7] phonologists have been drawn, moth-like, toward the beckoning flame of psychological reality. Much of contemporary phonological theory, to the extent that it seeks to model "competence," has at least a pretension to some form of psychological explanation. Some work presents this aim explicitly (e.g., Chomsky and Halle 1968, who from the beginning speak of grammar, their object of investigation, as a "system of rules represented in the mind of the speaker-hearer" [1]; or Prince and Smolensky 1993, chapter 10, who situate Optimality Theory relative to other approaches to the cognitive modeling of language, such as connectionism). Other work, sometimes ostensibly in the same tradition, either downplays the connection to cognition (Hyman 2018b, on "what's in the language" versus "what's in the head" [597]) or actively repudiates it (Ladefoged 2005 on language "as a social institution rather than a mental concept" [1]). Often, though, and the attentive reader will no doubt have begun to recognize a theme here, it is not always clear where a given analysis or model stands in this regard, inviting others to make assumptions that aren't always warranted. As with the other questions we have addressed, there is little sense in criticizing a model for being cognitively implausible if it makes no pretense to model cognition in the first place. One could always argue that a decent model should in fact attempt to model cognition, but again, this is a complaint of a very different nature.

At its inception, the Fujisaki model was not focused, directly at least, on questions of cognition. At the same time, to the extent that the phrase and accent components are characterized as modeling sets of "neuromotor commands carrying linguistic information" (Fujisaki, Ohno, and Wang 1998, 1), it is clear that these elements are meant to be "meaningful both linguistically and cognitively" (Fujisaki, personal communication).

I.6 Question 6: Transcription

A recurring question in the development of transcription systems, for either segmental or prosodic purposes, is whether to attempt to capture the contrastive categories of the language that underlie the speech signal, or to focus on the wide range of ways in which each category can be appropriately realized. With regard to segmental transcription, it is surely no exaggeration to claim that the creation and propagation of the International Phonetic Alphabet in the late nineteenth and early twentieth centuries laid groundwork without which we might never have seen the rapid progress in phonological thought that followed closely on its heels, and continues to the present day. There is a basic sense in which it is necessary to achieve (or at least approach) consensus as to *what* has happened in the course of some natural phenomenon before we can set about formulating explanations as to *why*, or *how*, and the IPA provides a way of doing that for spoken utterances. Thus, it has practical value to the extent that it provides us with a set of labels that, at least in principle, should be universally applicable and interpretable for the characterization of spoken language data, regardless of the language or context in which the speech was produced, and this was obviously and rightly one of its developers' central goals as well. Perhaps less obviously, and more importantly from a theoretical point of view, it also served to cement, behind those labels, an expanding set of putatively universal phonetic categories that should recur with greater or lesser regularity across the languages of the world. The idea is that we ought to be able to agree, for example, on whether or not a given speaker, in a given utterance, has produced an instance of voiceless, unaspirated, bilabial stop, and that we should be able to do so regardless of the phonological analysis we might like to give to that event down the road. Whether it represents an allophone of the phoneme /b/, or of /p/, or of something else again, depends on the language; the morphophonology of the lexical item in question; and so forth. That it was, regardless, an instance of phonetic [p], however, should in principle be the unimpeachable bedrock upon which any analysis must eventually stand.

The notion of universal phonetic categories has its problems, of course, and its detractors have not been shy in saying so. While parts of this universal inventory have seemed relatively uncontroversial, in other parts it has been less clear whether or where to draw boundaries between would-be categories, and whether to dignify particular candidates with their own symbols, or to relegate them instead to expression via diacritics draped around more respectable categories. Varieties of obstruents occurring somewhere around the alveopalatal (or alveolopalatal, or palato-alveolar, or postalveolar, or prepalatal) places of articulation, with or without concomitant tongue posture specifications, spring to mind.

The principle designed to mitigate the potential for unfettered expansion of the system of phonetic categories is that for a new symbol to be introduced, with all its ontological entailments, the sound in question must be contrastive in some language.[8] This is intended to guard against the adoption of distinct symbols to represent sound characteristics that are judged to be universally a matter of subphonemic detail rather than the sole feature minimally distinguishing one sound from another. In practice, though, it can be difficult in many cases to decide how such a principle should be applied.[9] Often the question demanding an answer seems to be, "Contrastive with what?" For example, one area in which the current state of the official IPA chart suggests a breakdown of application of this principle is in the profusion of variant symbols for degrees of nuance in the mid-central portion of the vowel space. The IPA chart currently contains six distinct symbols (excluding vowels such as [ɨ], [ɤ], or [ʌ], which are officially high or back, despite their occasional use as symbols for things lower or fronter than that seems to imply). These are [ɘ, ɵ, ə, ɜ, ɞ, ɐ]. In principle, these are all distinct vowel categories. We know what they are meant to sound like, and if we are ever in doubt, recordings exist of consensus productions to remind us. This is not the place to rehearse the individual careers of all these symbols.[10] No doubt it is the case that for each of them there exists, in some language somewhere, a mid-central vowel whose canonical realization sounds more like that one symbol than it does like any of the others. And presumably, that vowel is "contrastive." But with what? Most vowel systems that we know of have one mid-central vowel, if they have any. Some have two (assuming one of them is not better considered high, or back, or front, or what have you).

Obviously, no system has all of these vowels in contrast with one another. Indeed, it seems unlikely to us that even each potential two-way contrast implied by this set of symbols is convincingly represented in some attested language. But if it feels worrisomely like we have embarked on an effort to populate what is effectively a continuous space with discrete symbols whose optimal, final number is not specified by any existing principle, we might nonetheless take comfort in the thought that, with all these symbols, we're at least not likely to be missing anything.[11] For any given real-world mid-central vowel we might encounter, in other words, chances are we have a symbol that does it justice. Yet even so, the fussier of us remain dissatisfied (e.g., Barnes 2006), feeling compelled to note pedantically, for example, that while the IPA standard transcription for the Bulgarian vowel spelled "ъ" and traditionally romanized with "ă" is [ɤ] (Ternes and Vladimirova-Buhtz 1990), in fact the vowel in question sounds lower, and fronter, than that symbol implies, if, however, not so low and so front as to make any other IPA symbol more appropriate than that one. We have achieved, in other words, the worst of both worlds: a system at once with too much and with too little nuance built into it.[12]

This lengthy preamble is only to underscore the seriousness of the fact that, whatever uncertainty we might encounter with the transcription of segmental phonology, the situation with respect to the transcription of prosody is in every respect much, much worse. If there are parts of the vowel space in which natural boundaries between universal phonetic categories seem elusive, many, if not most, dimensions of segmental contrast seem reassuringly stable (at least at first glance). In prosody research, though, our abiding fear is that in the end, it may just be mid-central vowels all the way down. In thinking about the transcription of contrasting intonational pitch accents (or, for that matter, lexical tones), it seems much less obvious that a system of "universal phonetic categories," ripe for precise transcription, is waiting to be discovered. Even the contrast principle is of less use here, insofar as we struggle in prosody to produce convincing analogues of the lexical minimal pairs that make contrast judgments possible in segmental phonology. What kinds of things an International Prosodic Alphabet should aspire to encode, and indeed, whether such an alphabet is worth creating to begin with, are questions that at the moment of writing, elicit impassioned disagreement from otherwise even-tempered and well-adjusted researchers.

The question of how these phonetic categories in any case might connect to the contrastive phonological categories of the language has profound implications for a transcription system. An example of this issue can be found in the history of the development of the ToBI system, which took place over a series of four meetings in the early 1990s, first convened by Victor Zue of MIT. The intention was to develop a consensus transcription system for prosody, drawing on two sets of recently available resources: the emerging theory of the prosodic hierarchy, to be integrated with the emerging theory of intonation, and the accumulation of large databases of relatively spontaneous speech. These databases required hand-labeling of prosodic structure to enable machine learning of the signal characteristics associated with that structure, and there were a number of obstacles to the practical attainment of this goal. First, hand labeling is shockingly slow and costly, making it advantageous to pool labeled data, but second, different laboratories could not easily share their labeled data to form larger resources. That was because, over the previous half-century, different laboratories had developed idiosyncratic prosodic transcription systems, which did not allow for convenient sharing of labeled data. The goal of the workshops was to address this problem by developing a consensus system for prosodic transcription, so that laboriously labeled data could be more easily shared across laboratories and platforms.

During these developmental meetings, with a roomful of twenty to thirty meeting participants, all actively working on prosody or intonation but from widely different perspectives and with strikingly different goals, the discussion was lively, to say the least. One particularly striking division emerged between those who felt that the proposed system should capture all and only the prosodic aspects that were discernible in the acoustic signal of a given utterance, and those who felt that it should reflect the phonological contrasts in the prosody of the language, whether or not they were directly measurable in the signal, thereby enabling later analysis of the range of ways in which a given category can be signaled. In the absence of certainty about what the contrastive categories of English prosody are, those in the first camp were loath to label anything beyond what they saw in the signal, while those in the second camp were concerned that a purely surface approach to labeling would miss the point of the enterprise.

At times during the discussion, it seemed that only a thin veneer of civilized convention prevented the spilling of blood on the floor. The need for some kind of consensus was strong, however, and the outcome of these meetings, the ToBI transcription system, was a compromise. Some predictable aspects of the then-current phonology were eliminated in favor of capturing surface variation explicitly (e.g., explicit marking of downstep with the symbol !, rather than assuming downstepping of an H* pitch accent after a preceding bitonal accent), and some surface aspects of the acoustic signal were eliminated from explicit marking (e.g., declination across an intonational phrase, and relative scaling of accent-related F0 between phrases). The tension about the extent to which phonological analysis was baked into the system still remains problematic for many users and has led to proposals for alternative transcription systems that hew more closely to the surface form of the signal (e.g., RaP, and the ongoing effort to develop a phonetic-level IPrA).

As an aside, we note that a potential solution to this problem is to view prosodic transcription as a matter of specifying the set of individual acoustic-phonetic cues[13] to the prosodic structure of an utterance. For example, a number of different cues to the presence of a prosodic boundary have been observed, including final lengthening, F0 excursion (edge tone), silent pause, F0 reset (up or down), and glottalization (initial or final in the word), and some preliminary reports suggest that these individual cues are processed by listeners. Brugos et al. (2018) have reported that the greater the number of cues present at a boundary, as identified by human labelers annotating individual cues, the more likely an independent set of ToBI labelers are to mark an intonational phrase boundary at that location. Along the same lines, Brugos et al. (2019) proposed a system for labeling individual correlates of prosodic disfluencies, based on work by Arbisi-Kelm (2006).

For purposes of symmetry with our presentation of the preceding questions, we should note that the question of prosodic transcription has not been of central concern to practitioners of the Fujisaki model. The model indeed is relevant to questions of transcription only in the sense that it makes assumptions about the kinds of things that are relevant phonologically in prosodic systems, and hence conceivably suggests that certain elements require transcription norms. Fujisaki and colleagues, to our knowledge, takes no official position on how best to transcribe prosody, and even suggests at one point (Mixdorff and Fujisaki 2000) that something like ToBI could serve as a system of category labels that would be interpreted phonetically in terms of the Fujisaki model.

As must be obvious by now, the purpose of the six questions we posed to our chapter authors was to reveal the goals of each of the major approaches to prosody that have been proposed over the past few decades. We hope that by highlighting the similarities and differences in the aims of each theory, the individual chapters will make it easier for practitioners to determine which approach best suits their needs. We leave it to our readers to decide the extent to which each of the following chapters achieves the level of explicitness that we believe it behooves us all to aspire to.

Notes

1. This is not to suggest the literal transfer of a single set of invariant motor or perceptual patterns from one word to another. Rather, each “reuse” of a category would be modulated by the full range of systematic, complex, contextual variability that has been documented in natural speech. The point is that all these patterns of variability, however expansive and multifaceted, would connect ultimately to the categories, as uttered in a particular pragmatic, prosodic, or other context rather than directly to a particular lexical item or semantic element. It is this intermediate connection that makes the problem tractable for language users.

2. This example, incidentally, also represents what we hold to be the pinnacle of the prosodist’s refined art of composing lengthy carrier sentences out of nothing but sonorant segments. We salute its creator accordingly.

3. In the Fujisaki model, superposition, like everything else, is a matter of phonetic implementation. In autosegmental phonology, the process of tier conflation (McCarthy 1986) was hypothesized to be carried out during the phonological derivation and, potentially, could be extrinsically ordered relative to other phonological processes.

4. There are in fact certain tone strings that are often treated, informally at least, as though it were the entire contour with which utterance-level meanings were associated, making the contour more like the morpheme, and the tones that comprise it then more like phonemes—for example, the “contradiction contour,” the “surprise-redundancy contour” (Sag and Liberman 1975), and the “Eastern European question contour” (Ladd [1996] 2008).

To the extent that the meanings of these contours are not built compositionally from their constituent tones, they would have a status something like that of phrasal idioms in syntax and semantics.

5. Actually, one persistent critique of the AM model (e.g., Mixdorff 2002) is that it does not in fact have an agreed-on, explicitly stated encoding function to begin with, and thus is difficult to evaluate in these terms. In any case, the example we are considering here is an entirely hypothetical one.

6. In fairness, it is also possible that most phonologists at the time had simply never heard of diphone synthesis, and, indeed, various proposals reminiscent of di- or triphones have in fact surfaced in the phonological literature (e.g., the tonal transitions of Dilley [2005] or the quantized contours of Q theory [Shih and Inkelas 2019]). Arguments for those proposals, however, were largely devoid of references to synthesis.

7. See, e.g., Jakobson's 1942 criticism of the "naive psychologism" of Baudouin de Courtenay and Saussure (Jakobson 1990, 224–230).

8. See the historic Principle 1 of the International Phonetic Association, first promulgated in 1888: "There should be a separate letter for each distinctive sound; that is, for each sound which, being used instead of another, in the same language, can change the meaning of a word" (Abramson 1988).

9. "It was quite clear to everybody at the Kiel convention that this is not the best possible chart; but it was also clear that we could not agree on how it should be improved" (Ladefoged 1990, on the discussions leading to the 1989 Kiel revisions to the IPA).

10. Catford (1990) played a central role in the enrichment of this portion of the vowel chart, which was officially carried out in 1993. Interestingly, at that time, one of these symbols was accidentally adopted in mirror image, owing to a typographical error. In 1996, this mistake was corrected, with the erroneous "closed epsilon," [ɘ], banished in favor of the intended "closed reversed epsilon," [ɞ]. Catford himself apparently preferred "barred open O" for this purpose, but was overruled for reasons that are not known to us (Pullum and Ladusaw 1996, 1998).

11. This is perhaps somewhat uncharitable. The substance of Catford's argument in favor of the proliferation of mid-central vowel qualities relied on a principle of parallelism within the system of Cardinal Vowels. If there are four total heights and three "horizontal" locations for vowel targets in this system, and with rounding distinctions separating primary and secondary versions of each, then we need at least four symbols to represent central vowels at heights two and three (open-mid and close-mid), in both round and unround variants. As a practical argument, however, he cites the need to distinguish "variants of the vowel in *bird*" at least at open-mid, mid, and close-mid levels. Note that any need to distinguish these things symbolically in transcription is not based on principles of contrast.

12. In some cases, what seems to happen is that transcribers fail to find a symbol that adequately reflects the sound in question, but they worry that choosing the closest option implies a level of precision and nuance that is unwarranted, leading them to opt out of IPA altogether, as when phonologists render Korean "tense" stops with an asterisk [such as p*], effectively meaning "this sound, in this language, whatever it actually is."

13. We use the term *cues* informally here, recognizing that some correlates of linguistic categories that are present in the signal may not be actively used by listeners or explicitly planned by speakers. Determining which correlates actually function in this way requires extensive experimentation, only some of which has been carried out to date.

References

Abramson, Arthur. 1988. "The Principles on Which the IPA Should Be Based." *Journal of the International Phonetic Association* 18 (2): 66–68.

Arbisi-Kelm, T. 2006. "An Intonational Analysis of Disfluency Patterns in Stuttering." PhD diss., UCLA.

Barnes, J. 2006. *Strength and Weakness at the Interface: Positional Neutralization in Phonetics and Phonology*. Berlin: Mouton de Gruyter.

Bell, C. C., H. Fujisaki, J. M. Heinz, and K. N. Stevens. 1961. "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques." *Journal of the Acoustical Society of America* 33 (12): 1725–1736.

Brugos, A., M. Breen, N. Veilleux, J. Barnes, and S. Shattuck-Hufnagel. 2018. "Cue-Based Annotation and Analysis of Prosodic Boundary Events." In *Proceedings of the 9th International Conference on Speech Prosody*, June 13–16, 2018, Poznań, Poland, edited by Katarzyna Klessa, Jolanta Bachan, Agnieszka Wagner, Maciej Karpiński, and Daniel Śledziński, 245–249.

Brugos, A., A. Langston, S. Shattuck-Hufnagel, and N. Veilleux. 2019. "A Cue-Based Approach to Prosodic Disfluency Annotation." In *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia, edited by Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren, 3413–3417.

Catford, J. C. 1990. "A Proposal Concerning Central Vowels." *Journal of the International Phonetic Association* 20:26–28.

Cheney, D. L., and R. M. Seyfarth. 1990. *How Monkeys See the World: Inside the Mind of Another Species*. Chicago: University of Chicago Press.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper.

Clements, G. N. 1991. "Place of Articulation in Consonants and Vowels: A Unified Theory." *Working Papers of the Cornell Phonetics Laboratory* 5:77–123.

Dilley, Laura C. 2005. "The Phonetics and Phonology of Tonal Systems." PhD diss., MIT.

Fant, Gunnar. 1960. *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton.

Fujisaki, H. 1981. "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing: Acoustical Analysis and Physiological Interpretations." In *Proceedings of the Fourth F.A.S.E. Symposium on Acoustics and Speech*, 2:57–70.

Fujisaki, H. 1988. "A Note on the Physiological and Physical Basis for the Phrase and Accent Components in the Voice Fundamental Frequency Contour." In *Vocal Fold Physiology: Voice Production, Mechanisms and Functions: Proceedings of the Fourth F.A.S.E. Symposium on Acoustics and Speech55*, edited by O. Fujimura, 347–355. New York: Raven.

Fujisaki, H. 1997. "Prosody, Models, and Spontaneous Speech." In *Computing Prosody*, edited by Y. Sagisaka, N. Campbell, and N. Higuchi, 27–42. Berlin: Springer-Verlag.

Fujisaki, H. 2004. "Information, Prosody, and Modeling—With Emphasis on Tonal Features of Speech." In *Speech Prosody 2004*, International Conference, Nara, Japan, March 23–26, 2004, edited by Bernard Bel and Isabelle Marlien, 1–10.

Fujisaki, H., and K. Hirose. 1982. "Modelling the Dynamic Characteristics of Voice Fundamental Frequency with Applications to Analysis and Synthesis of Intonation." *Preprints of Papers from the Working Group on Intonation, Thirteenth International Congress Linguists*, Tokyo, Japan, edited by Hiroya Fujisaki and Eva Gårding, 57–70.

Fujisaki, H., and K. Hirose. 1984. "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese." *Journal of the Acoustical Society of Japan (E)* 5 (4): 233–242.

Fujisaki, H., and S. Nagashima. 1969. "A Model for the Synthesis of Pitch Contours of Connected Speech." *Annual Report of the Engineering Research Institute, University of Tokyo* 28:53–60.

Fujisaki, H., S. Ohno, S., and C. Wang. 1998. "A Command–Response Model for F0 Contour Generation in Multilingual Speech Synthesis." In *Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis*, 299–304.

Fujisaki, H., C. Wang, S. Ohno, and W. Gu. 2005. "Analysis and Synthesis of Fundamental Frequency Contours of Standard Chinese Using the Command-Response Model." *Speech Communication* 47:59–70.

Gårding, E. 1983. "A Generative Model of Intonation." In *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd, 11–25. Berlin: Springer.

Glewwe, E., J. Zymet, J. Adams, R. Jacobson, A. Yates, A. Zeng, and R. Daland. 2018. "Substantive Bias and the Acquisition of Final (De)Voicing Patterns." Paper presented at the 92nd Annual Meeting of the Linguistic Society of America, Salt Lake City, Utah, January 4–7.

Gordon, Matthew. 2016. *Phonological Typology*. Oxford: Oxford University Press.

Grice, Martine, D. Robert Ladd, and Amalia Arvaniti. 2000. "On the Place of 'Phrase Accents' in Intonational Phonology." *Phonology* 17:143–185.

Halle, Morris, and Kenneth Stevens. 1959. "Analysis by Synthesis." In *Proceedings of the Seminar on Speech Compression and Processing*, vol. 2, paper D-7, edited by W. Wathen-Dunn and L. E. Woods.

Hirst, D. 2011. "The Analysis by Synthesis of Speech Melody: From Data to Models." *Journal of Speech Sciences* 1 (1): 55–83.

Hockett, Charles F. 1958. *A Course in Modern Linguistics*. New York: Macmillan.

Hyman, Larry M. 2018a. "What Is Phonological Typology?" In *Phonological Typology*, edited by Larry M. Hyman and Frans Plank, 1–20. Berlin: Mouton de Gruyter.

Hyman, Larry M. 2018b. "Why Underlying Representations?" *Journal of Linguistics* 54 (3): 591–610.

Jakobson, Roman. 1990. "The Concept of the Phoneme." In *On Language*, edited by Linda R. Waugh and Monique Monville-Burston, 217–242. Cambridge, MA: Harvard University Press.

Kiparsky, Paul. 2006. "The Amphichronic Program vs. Evolutionary Phonology." *Theoretical Linguistics* 32 (2): 217–236.

Ladd, D. Robert. (1996) 2008. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Ladd, D. Robert. 2014. *Simultaneous Structure in Phonology*. Oxford: Oxford University Press.

Ladefoged, Peter. 1990. "The Revised International Phonetic Alphabet." *Language* 66 (3): 550–552.

Ladefoged, Peter. 2005. "Features and Parameters for Different Purposes." *UCLA Working Papers in Phonetics* 115 (104): 1–13.

Liu, F, Y. Xu, S. Prom-on, and A. C. L. Yu. 2013. "Morpheme-like Prosodic Functions: Evidence from Acoustic Analysis and Computational Modeling." *Journal of Speech Sciences* 3:85–140.

Martinet, André. 1949. "La double articulation linguistique." *Travaux du Cercle Linguistique de Copenhague* 5:30–37.

McCarthy, John. 1986. "OCP Effects: Gemination and Antigemination." *Linguistic Inquiry* 17 (2): 207–263.

Mixdorff, Hansjörg. 1998. "Intonation Patterns of German—Model-based Quantitative Analysis and Synthesis of f0-Contours." PhD diss., Technische Universität Dresden.

Mixdorff, Hansjörg. 2002. "Speech Technology, ToBI and Making Sense of Prosody." In *Speech Prosody 2002, International Conference*, Aix-en-Provence, France, April 11–13, 2002; ISCA Archive. http://www.isca-speech.org/archive/sp2002.

Mixdorff, Hansjörg, and Hiroya Fujisaki. 2000. "Symbolic versus Quantitative Descriptions of F0 Contours in German: Quantitative Modelling Can Provide Both." In *Prosody 2000: Speech Recognition and Synthesis*, Krakow, Poland, October 2–5.

Moreton, Elliott. 2008. "Analytic Bias and Phonological Typology." *Phonology* 25 (1): 83–127.

Moreton, Elliott, and Joe Pater. 2012. "Structure and Substance in Artificial-Phonology Learning, Part II: Substance." *Language and Linguistics Compass* 6 (11): 702–718.

Öhman, S., and J. Lindqvist. 1965. "Analysis-by-Synthesis of Prosodic Pitch Contours." *Speech Transmission Laboratory: Quarterly Progress and Status Reports* 6 (4): 1–6.

Pierrehumbert, J. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pierrehumbert, Janet, and Julia Hirschberg. 1990. "The Meaning of Intonational Contours in the Interpretation of Discourse." In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Prince, Alan, and Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report 2. New Brunswick, NJ: Rutgers University Center for Cognitive Science.

Pullum, Geoffrey K., and William A. Ladusaw. 1996. *Phonetic Symbol Guide*. Chicago: University of Chicago Press.

Pullum, Geoffrey K., and William A. Ladusaw. 1998. "Vowel Charts and Central Vowel Transcriptions in American and IPA Traditions." *Publication of the American Dialect Society* 80 (1): 5–33.

Sag, Ivan, and Mark Liberman. 1975. "The Intonational Disambiguation of Indirect Speech Acts." In *Papers from the Eleventh Regional Meeting, Chicago Linguistic Society*, edited by Robin E. Grossman, L. James San, and Timothy J. Vance, 487–497.

Sandler, W., M. Aronoff, I. Meir, and C. Padden. 2011. "The Gradual Emergence of Phonological Form in a New Language." *Natural Language and Linguistic Theory* 29:503–543.

Sapir, E. 1933. "The Psychological Reality of Phonemes." In *Selected Writings of Edward Sapir in Language, Culture and Personality*, edited by D. G. Mandelbaum, 46–60. Berkeley: University of California Press.

Shih, Stephanie, and Sharon Inkelas. 2019. "Autosegmental Aims in Surface-Optimizing Phonology." *Linguistic Inquiry* 50 (1): 137–196.

Ternes, E., and T. Vladimirova-Buhtz. 1990. "Bulgarian." *Journal of the International Phonetic Association* 20 (1): 45–47.

Yip, M. J. 1980. "The Tonal Phonology of Chinese." PhD diss., MIT.

1 The Autosegmental-Metrical Model of Intonational Phonology

Amalia Arvaniti

1.1 Introduction

Intonation refers to the linguistically structured and pragmatically meaningful modulation of F0. It is used in all languages and is specified at the phrasal level by means of a complex interplay of metrical structure, prosodic phrasing, syntax, and pragmatics; these factors determine where pitch movements will occur and of what type they will be. Although the main phonetic exponent of intonation is F0, the two are not synonymous, as F0 is used for purposes other than intonation, including the encoding of lexical tonal contrasts (as in Mandarin or Cantonese) and the signaling of paralinguistic information, such as boredom, anger, or excitement. Intonation is seen here as part of a language's *prosody*, an umbrella term that encompasses interacting phenomena that include rhythmic structure, prominence, and prosodic phrasing. The term *prosody* is preferred over the more frequent *suprasegmentals*, so as to avoid the layering metaphor inherent in the latter (cf. Beckman and Venditti 2011). This is so because prosody is not supplementary to segmentals (as the layering metaphor implies); it is an integral part of speech organization and intricately linked to the realization of segments (e.g., Beckman and Edwards 1994; Baltazani 2006; Niebuhr 2008, 2012).

The autosegmental-metrical model of intonational phonology is a model of intonational structure developed over the past forty years on the basis of empirical evidence from speech production and perception. The essential tenets of the model are largely based on Janet Pierrehumbert's (1980) dissertation, with additional refinements built on experimental research in a large number of languages (see Gussenhoven 2004; Jun 2005c, 2014b for surveys). The term *autosegmental-metrical* (henceforth AM), coined by Ladd (1996), reflects the connection between two subsystems of phonology required for intonation, an autosegmental tier representing intonation's melodic part, and metrical structure representing prominence and phrasing.

AM makes a principled distinction between the phonology of intonation and its phonetic realization. This distinction allows it to adapt to the descriptive needs and analytical requirements of prosodically varied languages, rather than be tethered to a conception that best fits a particular prosodic type (cf. Fujisaki 2004; Xu 2005; Xu et al. 2015). Crucial to AM's success has been the central role it gives to the compositionality of tunes. Specifically, AM recognizes that one of the major issues in modeling intonation is how to analyze continuous (and often far from smooth) pitch contours into a series of primitives. This is an essential step because the melodic component of intonation is postlexical, and thus primitives are not as readily identifiable as the tones of a tone language, such as Mandarin. Not all models of intonation fully engage with compositionality in that they often treat contours as the primitives that have to be faithfully modeled, an approach that leads to problems in both analysis and modeling as discussed in some detail in section 1.3.4 (see also Arvaniti and Ladd 2009; Arvaniti 2011).

To illustrate this essential point consider the pitch contours in figure 1.1. This figure shows melodies from three languages: English, Greek, and Polish. In each case, the utterance on the left is monosyllabic, while the one on the right contains several, often polysyllabic words. The contours of the two utterances are similar to each other but are not identical: neither can be said to be a stretched-out or squeezed version of the other. This can be seen if we compare across languages: while all short contours are rise-fall-rises, the longer contours are not as similar to each other (or to their short counterpart). This is a first indication that the contours are composed of a number of elements, each with its own properties. Despite these differences, in all three cases both contours are recognized by native speakers as realizations of the same melody in terms of form and pragmatic function: the two English contours imply incredulity (Ward and Hirschberg 1985; Hirschberg and Ward 1992); the Greek contours are used by default with straightforward *wh*-questions (Arvaniti and Ladd 2009; Baltazani, Gryllia, and Arvaniti 2020); the two Polish contours are instances of the melody used to call a person for a routine reason (Arvaniti, Żygis, and Jaskula 2017). As a great deal of research has shown, neither the differences in form nor the perceived similarity in each pair of contours is accidental (among many, Redi 2003; Prieto 2005; Prieto and Torreira 2007; Arvaniti and Ladd 2009). Rather, contours exhibit what has been termed "lawful variability" by Arvaniti and Ladd (2009), that is, variation that is context-dependent and systematic, involving variables such as length of utterance (as shown in figure 1.1) and the position of stressed syllables (see Arvaniti 2016, for a detailed discussion of additional sources of lawful variability). The role of stress is illustrated in figure 1.2, which shows the "incredulity" contour from figure 1.1 used with names with different stress patterns. Primary stress is on the first syllable of "Mary" but on the second syllable of "Michelle"; the effect of this difference is evident in the position of the peak, which aligns with the first syllable of "Mary" but the second syllable of "Michelle"; as a result, the last syllable of "Mary" is rising while that of "Michelle" is rising-falling-rising instead.

Given data such as these, it is clear that any successful model of intonation should, at a minimum, account for two things. First, it should account for the differences observed among the contours of each language, such as those in figures 1.1 and 1.2. Second, it should provide a generalization capturing the fact that speakers of that language perceive these contours as instantiations of the same melody. As demonstrated in Arvaniti and Ladd (2009), these goals can be achieved only by abstracting away from phonetic detail and overall contour shape to a phonological structure that reflects the commonalities between instantiations and provides a principled way of explaining their differences. Doing so means that posited abstract representations are predictive: they allow one to extrapolate how a given melody relates to the segmental string, and thus they provide a blueprint for how to use the same melody with utterances of varying lengths and prosodic structures. Predictions also mean that AM analyses can be falsified using data that test whether the predicted effects of lawful variability are actually observed in speech. Finally, a successful model should be able to address the first two points in a variety of languages with distinct prosodic properties; this requires flexible primitives and a recognition of differences between prosodic types. Achieving these aims and testing them experimentally are the central goals of AM.

Everything discussed thus far can be juxtaposed to other models of intonation, which are focused instead on faithfully reproducing pitch contours, such as Xu's PENTA (e.g., Xu 2005) and the Fujisaki model (e.g., Fujisaki 2004). Such models can be very successful in modeling F0 but not necessarily in modeling intonation as they cannot provide principled

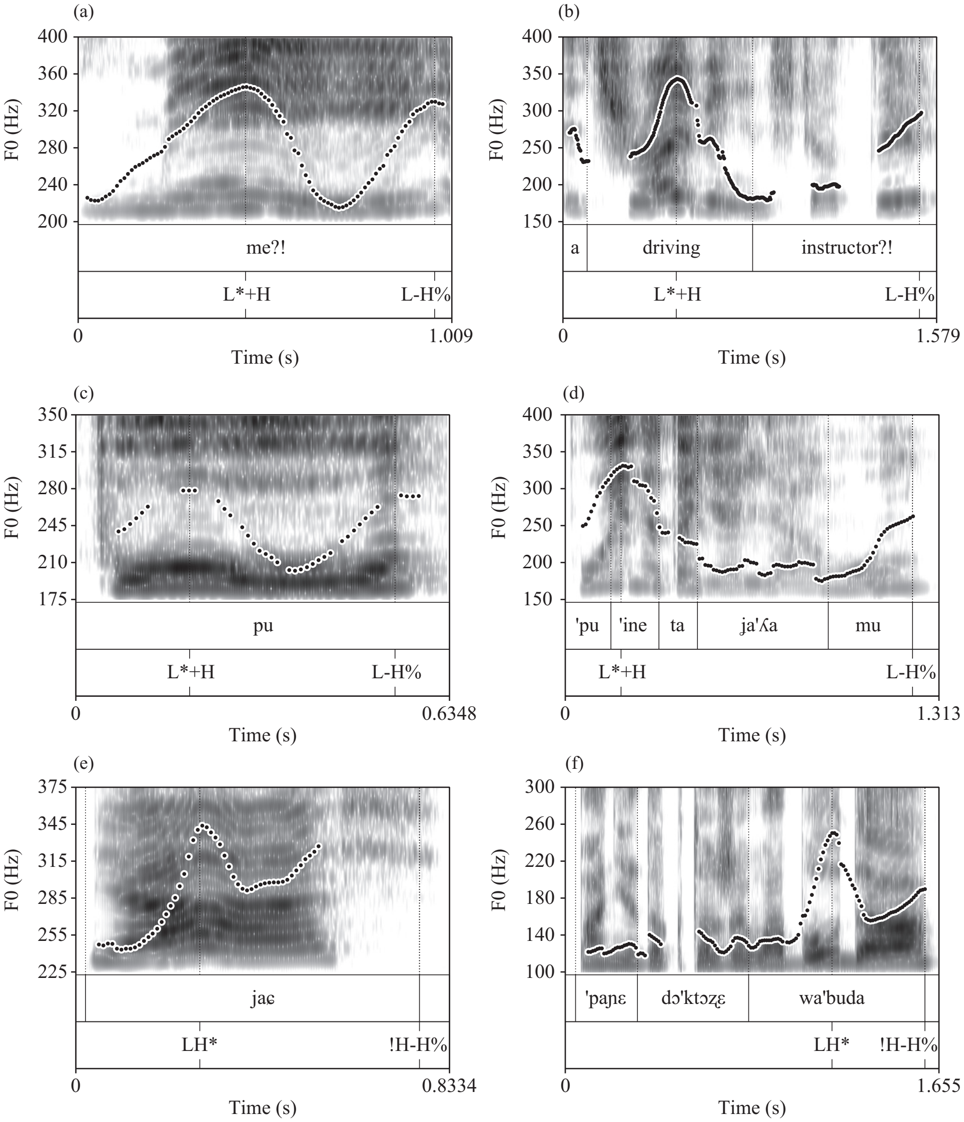


Figure 1.1

Spectrograms and F0 contours in three languages illustrating the same melody as realized on a monosyllable and a longer utterance: In English (top), using the "surprise-redundancy" contour (e.g., Hirschberg and Ward 1992); in Greek (middle), using the default melody for *wh*-questions ((a) 'where'; (b) 'where are my glasses'; Arvaniti and Ladd 2009); in Polish (bottom), using the default calling melody ((a) *Jaś* (name), (b) *Panie Doktorze Łabuda* 'Mr Dr Wabuda'); adapted from Arvaniti, Żygis, and Jaskuła 2017.

descriptions of melodies, only model them a posteriori. Furthermore, these models cannot provide any principled explanations for differences observed between realizations of the same melody, nor, as it turns out, can they better model certain types of lawful variability (see section 1.3.4; for a thorough discussion, see Arvaniti and Ladd 2009, 2015).

The main tenets and theoretical principles of AM are discussed in section 1.2; phonetic realization is discussed in section 1.3; section 1.4 discusses the treatment

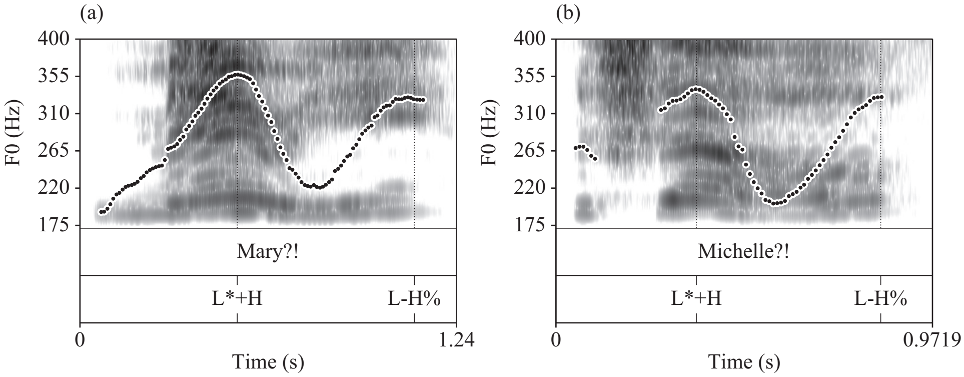


Figure 1.2
Spectrograms and F0 contours of "Mary" (left) and "Michelle" (right) with the surprise-redundancy contour, illustrating differences in coordination between segments and contour depending on the position of the stressed syllable.

of intonational meaning in AM; section 1.5 discusses typology and section 1.6 covers applications of AM and the role of transcription.

1.2 AM Phonology

1.2.1 AM Essentials

According to the basic principles of AM, intonation is phonologically represented as a string of low (L) and high (H) tones and combinations thereof (Pierrehumbert 1980; Beckman and Pierrehumbert 1986; Ladd 2008). L and H tones are the abstract symbolic (i.e., phonological) primitives of intonation. Their identity as Hs and Ls is largely determined by phonetic observation and defined in relative terms: H is used to represent tones deemed to be high in a melody with respect to the speaker's range and other tones in the same contour; L is used to represent tones deemed to be low by the same criteria (cf. Pierrehumbert 1980, 68–75). Tonal entities (which may be composed of one or more tones, as shown in section 1.2.2) are morphemes with pragmatic meaning; all tonal entities in a melody contribute compositionally to the pragmatic interpretation of an utterance (Pierrehumbert and Hirschberg 1990).

Tones are *autosegments*, that is, they are independent of the segmental string. They are phonologically associated to structural positions in the metrical representation of an utterance. It is this connection that has given rise to the term *autosegmental-metrical*. In particular, tones associate either with phrasal boundaries or constituent heads. In the latter case, association is often with feet, so informally we can say that these tones associate with stressed syllables. These tones are called *pitch accents*; they are notated with a star, for example, H*. The last and most important of pitch accents in a phrase is referred to as the *nuclear* accent (see also section 1.2.5). Tones that associate with boundaries are collectively known as *edge tones*; many analyses posit two types of edge tones, *boundary tones* and *phrase accents*, notated with % and -, respectively (e.g., H%, H-); the main role of edge tones is to demarcate phrasal boundaries. All languages investigated so far have edge tones that associate with right boundaries; left edge boundary tones have also been posited for several languages, including English (Pierrehumbert 1980), Basque (Elordieta and Hualde 2014), and Mongolian (Karlsson 2014).

Phonetically, *tones* are said to be realized as tonal targets, that is, as specific points in the contour, while the rest of an F0 curve is derived by interpolation between these targets. *Tonal targets* are usually turning points such as peaks, troughs, and elbows in the contour; they are defined by their scaling and alignment. *Scaling* refers to the value of the targets in terms of F0 or pitch.[1] *Alignment* is defined as the position of a tonal target relative to a specific tone-bearing unit (TBU) with which that is meant to co-occur (Arvaniti, Ladd, and Mennen 1998, 2006a, 2006b; Arvaniti and Ladd 2009). The identity of TBUs varies by language, depending on syllable structure, but we can roughly equate TBUs with syllable nuclei. The TBUs with which tones phonetically co-occur are related to the metrical positions with which the tones associate in phonology: thus, pitch accents typically co-occur with stressed syllables; edge tones are realized on peripheral TBUs such as phrase-final vowels. The interpolation between targets is normally linear and can be conceived of as the equivalent of an articulator's trajectory between two constrictions (Browman and Goldstein 1992). However, Pierrehumbert (1981) also recognizes a type of sagging interpolation used between H* accents in English, such that F0 curves "sag" between two peaks, with sagging increasing with increased temporal distance between the peaks (see figure 1.3). I return to this point in section 1.3.3.

This understanding of intonational structure has important consequences for modeling. The formal separation between tones and TBUs (an insight we owe to analyses of tone in African tone languages, such as that of Goldsmith 1976), is that tones and TBUs are not always in a one-to-one relationship as they are in models like PENTA (Xu et al. 2015). In AM, utterances like those on the left column of figure 1.1 include more tones than TBUs, while the relationship is reversed in the utterances on the right, which have more TBUs than tones. This in turn means that while in some instances several tones are associated with the same TBU, many TBUs in others are phonologically unspecified for tone. This conceptualization has repercussions for phonetic implementation as discussed in section 1.3.4.

1.2.2 The Structure of Tonal Events

Ls and Hs can be tonal events in their own right but they can also combine into complex units. The most commonly found combinations are bitonal pitch accents; for example, L*+H- and L-+H*. This notation reflects Pierrehumbert's (1980) analysis according to which the starred tone is the stronger of the two and the only one that is phonologically associated with a constituent head. The weak tone that leads or trails

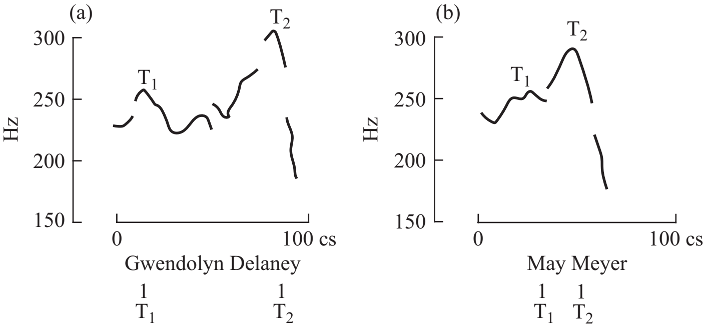


Figure 1.3
Examples of sagging interpolation with significant sag (left) and minimal sag (right). *Source*: Courtesy of Pierrehumbert (1981).

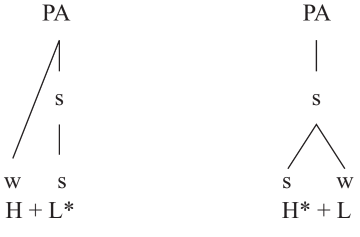
the starred tone is a floating autosegment (an autosegment without an association to a specific metrical position). Phonetically the weak tone precedes or follows the starred tone by a fixed amount of time but it can also spread.

Since Pierrehumbert, both the notations used for bitonal pitch accents and our understanding of their phonetics and phonology have evolved. The notation advocated by Pierrehumbert has largely been replaced by variants that dispense with either the dash for the weak tone or the dash and plus sign; thus, a bitonal accent represented in Pierrehumbert's system as L* + H- is now usually notated as L* + H or LH*. Although these notations have often been used interchangeably, research shows systematic differences between types of bitonal accents. Such differences prompted Jun and Fletcher (2014) to propose that the two most common notations, exemplified here by L* + H and L*H, be used to represent distinct types of bitonal accents: accents with a loose affiliation of their tonal elements, represented with the plus sign (e.g., L* + H), and accents showing a close connection between their tones, represented without it (e.g., L*H).

Evidence for different relations between the tones of bitonal accents has been reported for some time. According to Grice (1995, 215ff), leading and trailing tones in English differ in status: leading tones associate with the syllable preceding the accented one (if one is available), while trailing tones occur a fixed interval in "normalized time" after the starred tone. Grice posits hierarchical structures to represent the differences (see (1)).

Arvaniti, Ladd, and Mennen (1998) presented experimental evidence for the independence of tones in bitonal accents. They showed that Greek prenuclear rising pitch accents consist of a low and a high target, interpreted as reflexes of a L and H tone, respectively, that have stable alignment with distinct segmental landmarks rather than each other. Frota (2002) has further shown that this type of "loose" bitonal accent can coexist in a system with pitch accents that show a closer connection between tones. Frota compared the alignment of nuclear falling accents in European Portuguese declaratives with broad and narrow focus. She found that the fall of the broad focus H + L* accent shows a constant timing relationship between the H and L tone. On the other hand, in the fall of the narrow focus H* + L accent, the H and L are timed with respect to distinct segments rather than each other. Frota uses the hierarchical representations in (1), first proposed by Grice (1995), to capture the difference. A shorthand would be to use HL* and H* + L, respectively, as in Jun and Fletcher (2014).

(1)



Data like those discussed support the view that bitonal accents are of at least two types. Some consist of a metrically strong tone, preceded or followed by a weak tone; this is exemplified by the English L* + H (Arvaniti and Garding 2007) and the Portuguese H + L* (Frota 2002). Others form loose units, such as the Greek prenuclear L* + H pitch accent (Arvaniti, Ladd, and Mennen 1998) or the Portuguese H* + L (Frota 2002), with each tone showing alignment with a particular segmental landmark. The latter pattern appears to be prevalent in that it has also been reported for English L + H* (Dilley, Ladd, and Schepman 2005), Dutch L + H* (Schepman, Lickley, and Ladd 2006), and

Catalan L + H*, L+ <H*, L* + H, and H + L* (Prieto 2014). Additional types may be possible, for example, along the lines advocated by Grice (1995), but empirical validation of this point is, to this author's knowledge, currently absent.

The Greek data of Arvaniti, Ladd, and Mennen (1998) have prompted Arvaniti, Ladd, and Mennen (2000) to question the notion of starredness altogether because neither the L nor the H tone aligns with the stressed syllable. The Greek data clearly contravene a strict interpretation of phonological association whereby association entails phonetic alignment, that is, co-occurrence with an associated TBU. This objection is moot, however, if association is understood to involve relationships between tones and structural positions (as advocated by Pierrehumbert and Beckman 1988) rather than specific TBUs. If this formalization of association is accepted, then the phonetic co-occurrence of tones with specific TBUs is likely but not necessary. Furthermore, phonetic alignment need not be the sole or most important criterion for determining starredness in a bitonal accent; metrical strength or relative importance of a tone should also be considered. Alternatively, it is possible to use the star notation symbolically to indicate that the tonal entity in question is a pitch accent but without repercussions for the phonetic alignment of individual tones. This is the solution advocated by Ladd and Schepman (2003), who attribute the sagging transition between H* accents of English to an L tone and propose that H* be replaced by (L + H)*. This conceptualization of the pitch accent requires greater abstraction in autosegmental representations, a point to which I return in section 1.5 (for a detailed discussion, see Arvaniti 2016).

As noted, bitonal pitch accents are the most frequently used complex AM primitives. More complex accents have been occasionally proposed (e.g., tritonal pitch accents with internal structure; Grice 1995), but they have not been widely adopted. Bitonal accents, on the other hand, can be found in several analyses (among many, Ladd and Schepman 2003 for English; Grice, Baumann, and Benzmüller 2005 for German; Prieto 2014 for Catalan; Chahal and Hellmuth 2014 on Arabic; Gooden 2014 on Jamaican Creole). Most such analyses focus more on phonetic transparency, that is, on a faithful representation of the pitch movement involved, and not necessarily on the theoretical ramifications of such representations for starredness and the internal structure of tonal events. Despite the findings and proposals discussed, it is fair to say that within AM, it is typically assumed that one tone of a bitonal pitch accent is starred and co-occurs (to a greater or lesser extent) with the accented syllable.

Multitonal combinations are posited for edge tones as well; for example, Jun (2005a) posits boundary tones with up to five tones, such as LHLHL%, for Korean; Prieto (2014) posits a tritonal LHL% boundary tone for Catalan. It is important to note that the presence of multitonal edge tones is often dependent on other analytical decisions to do with the number of edge tones in a system and the division of labor between pitch accents and edge tones. Thus, in an analysis that includes phrase accents, a phrase-final rise with a low onset may be represented as L-H%, while in a system that includes only boundary tones, the same pitch movement may be represented as a LH% boundary tone. Alternatively, an analysis may incorporate the L tone into the nuclear accent; thus, a sequence that in one system would be analyzed as H* L-H% may be represented as H* + L H% in another (cf. the analyses of English intonation in Pierrehumbert 1980 and Gussenhoven 2004, chapter 15). Such differences rarely reflect disagreements regarding the identity of tones as Ls or Hs; rather, disagreements relate to the function of tones in the system and decisions as to how exactly to carve up a continuous F0 curve into component events. At present there are no strict and widely accepted criteria within AM that would allow researchers to decide in a principled way between analyses such as those discussed. Jun and Fletcher (2014) propose diagnostics that could help researchers determine whether phrase accents are necessary in the analysis of a given language; however, in their proposal phrase accents are predicated on the existence of intermediate phrases (see section 1.2.3).

Another solution to this problem has been presented by Ritter and Grice (2015), who used perception experiments to determine the composition of the accent in rise-fall contours in German. These contours have been analyzed as involving either a pitch accent with an onglide (L + H) or a pitch accent with an offglide (H + L), choices that have repercussions for the representation of the remainder of the contour. A different approach is taken by Gussenhoven (2016) who focuses on the repercussions of particular analytical decisions for the description of entire intonation systems. Gussenhoven considers two AM analyses of English intonation and argues that the one that includes phrase accents leads to both overanalysis and underanalysis of attested contours. Research along these lines, considering both experimental data and system-internal arguments, should help resolve such disputes regarding the analysis of other intonation systems as well.

1.2.3 Secondary Association

Secondary association is a mechanism that has been used for a variety of purposes in AM but has come to be strongly associated with phrase accents. Specifically, while the status and role of boundary tones are uncontroversial within AM, the status of phrase accents is not generally agreed on. Some AM systems dispense altogether with phrase accents either for reasons of parsimony or because they adopt a different conception of how the F0 contour is to be broken into constituents (see, e.g., Gussenhoven 2004, 2005 on Dutch; Frota 2014 on Portuguese, Gussenhoven 2016 on English). Others, such as Ladd (1996), have argued against phrase accents, considering them not a special category of tonal events but rather postnuclear pitch accents instead.

Among analyses that do posit phrase accents, complete consensus as to their nature is also lacking. In Pierrehumbert's (1980) analysis, which is inspired by Bruce's (1977) analysis of similar phenomena in Swedish, phrase accents are floating autosegments (akin to the weak tones of bitonal accents). Phonetically they are said to spread, covering the part of the utterance between the nuclear accent and the boundary tone. The phrase accent as a floating autosegment met with criticism in Ladd (1983). In response, Beckman and Pierrehumbert (1986) recast phrase accents as edge tones associated with the boundaries of intermediate phrases (ips), a phrasal constituent immediately below the intonational phrase (IP) in their prosodic hierarchy; for example, "Mel and Kate were late because they missed their bus" is likely to be uttered as two ips forming one IP: [[Mel and Kate were late]$_{\text{ip}}$ [because they missed their bus]$_{\text{ip}}$]$_{\text{IP}}$. Each level of phrasing is associated with a type of edge tone: ips with phrase accents and IPs with boundary tones. The proposal of Beckman and Pierrehumbert (1986), which links phrase accents to the ip, has been implicitly adopted by many analyses, including those for Greek (Arvaniti and Baltazani 2005), German (Grice, Baumann, and Benzmüller 2005), Jamaican Creole (Gooden 2014), and Lebanese and Egyptian Arabic (Chahal and Hellmuth 2014).

Pierrehumbert and Beckman (1988) first proposed the mechanism of secondary association to account for the fact that phrase accents often spread. Specifically, they proposed that edge tones may acquire additional links (i.e., secondary associations) either to a specific TBU or to another boundary. For example, they posit that English phrase accents are linked not only to the right edge of the ip as advocated in Beckman and Pierrehumbert (1986), but also to the left edge of the word carrying the nuclear pitch accent. This type of double association accounts for the fact that in English, phrase accents correspond to a stretch of the contour rather than to a single maximum or minimum. Similarly, in the analysis of Pierrehumbert and Beckman (1988), the Japanese phrasal H is associated with the left edge of an accentual phrase (AP, a constituent immediately below the ip in the Japanese prosodic hierarchy of Beckman and Pierrehumbert 1986), but also linked to a specific TBU within its AP. The general mechanism whereby edge tones have secondary associations to TBUs has also been used by Gussenhoven (2000) in his analysis of the intonation of Roermond Dutch, which assumes that boundary tones can be phonologically aligned both with the right edge of the phrase and with an additional leftmost position.

The analyses of Pierrehumbert and Beckman (1988) and Gussenhoven (2000) were the basis for a wider use of secondary association for phrase accents developed in Grice, Ladd, and Arvaniti (2000), who adopt the notion of the phrase accent but argue that the need for positing phrase accents in a given intonation system is orthogonal to the need for the ip (i.e., that phrase accents need not demarcate the right edges of ips). Grice, Ladd, and Arvaniti (2000) examined putative phrase accents in a variety of largely unrelated languages (Hungarian, Romanian, Greek, Cypriot, English, German, and Dutch). They showed that the phrase accents they examined are realized on either a peripheral TBU, as expected of edge tones, or an earlier TBU, often one that is metrically strong. Which of the two realizations prevails depends on whether the latter TBU is associated with another tone. This type of variation is illustrated in figure 1.4 with Greek polar questions, the melody of which is autosegmentally analyzed as L* H-L% (Grice, Ladd, and Arvaniti 2000; Arvaniti, Ladd, and Mennen 2006a). As can be seen, both the contour in the left panel and the contour on the right have a pitch peak close to the end of the utterance; this peak co-occurs with the stressed antepenult of the last word in the question in the left panel but with the last TBU in the question in the right panel. (Note that the stressed antepenult has low F0 in the right panel contour as it is the TBU on which the L* pitch accent is realized). Grice, Ladd, and Arvaniti (2000) attribute this difference in the alignment of the pitch peak to secondary association: the peak is the reflex of a H- phrase accent, which is associated with the phrasal boundary, but also has a *secondary association* to the last metrically strong syllable of the utterance. This

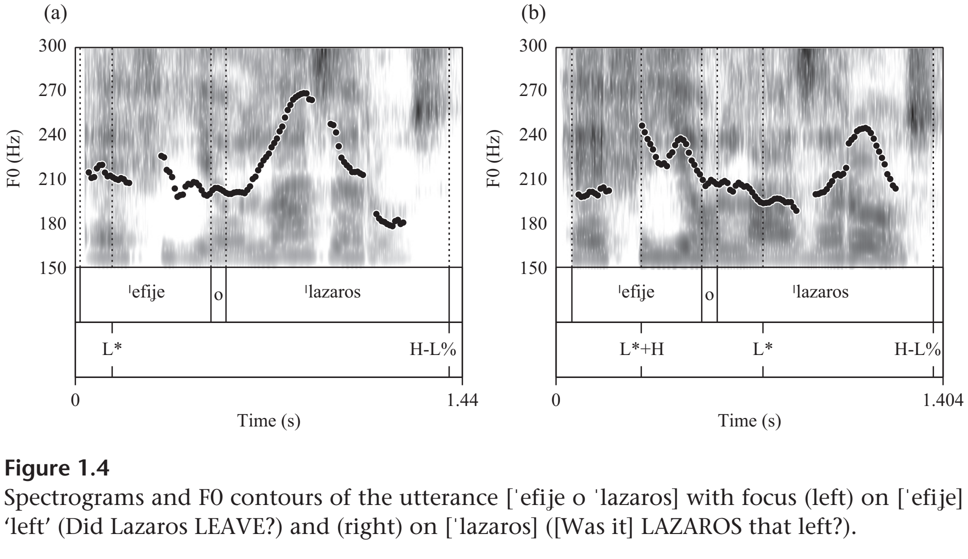
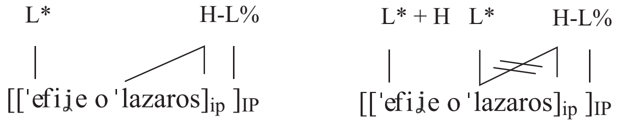


Figure 1.4
Spectrograms and F0 contours of the utterance [ˈefiʝe o ˈlazaros] with focus (left) on [ˈefiʝe] ‘left’ (Did Lazaros LEAVE?) and (right) on [ˈlazaros] ([Was it] LAZAROS that left?).

association is phonetically realized if this syllable is not associated with a pitch accent. The phonological structures involved are shown in (2).

(2) a. Did Lazaros LEAVE? b. [Was it] LAZAROS [that] left?



The data of Grice, Ladd, and Arvaniti (2000) partly vindicated Ladd (1996) by showing that some postnuclear tones do associate with metrically stressed syllables. Similar results have since been reported for Dutch (Lickley, Schepman, and Ladd 2005), English (Barnes et al. 2010), and Greek (Arvaniti and Ladd 2009), with all phrase accents showing "stress-seeking" behavior to some extent. However, Grice, Ladd, and Arvaniti (2000) also found that secondary associations are not always to metrically strong syllables; for example, in Hungarian, secondary association is with the penultimate syllable independently of stress. This suggests that phrase accents with secondary association cannot be equated with postnuclear pitch accents (but see Ladd 2008, chapter 8, for an alternative view). Arvaniti (2009a) has further shown that at least in Greek, postnuclear accents and phrase accents with secondary association to a metrically strong syllable are distinct. Postnuclear pitch accents of reduced scaling are optional and may appear in long polar questions when the nuclear L* pitch accent is early in the utterance; the H- phrase accent, on the other hand, is an obligatory element of the tune and is scaled much higher than the optional postnuclear pitch accents.

The mechanism of secondary association has also been used by Prieto, D'Imperio, and Gili Fivela (2005) to account for alignment differences observed in the realization of pitch accents in Catalan and in Neapolitan and Pisan Italian. Specifically, they posit that in addition to their primary association with a stressed syllable, pitch accents may have a secondary association to a metrical position, such as the first mora of the accented syllable or the right edge of the accented word. They argue that this is needed because in the linguistic systems they examined, pragmatic distinctions conveyed by pitch accents are such that tonal composition and starredness alone cannot capture the contrasts involved. Though the analysis of Prieto, D'Imperio, and Gili Fivela has not been used in subsequent work (for example, it is absent from Prieto's 2014 analysis of Catalan intonation), the issue of what to include in phonological representations and how abstract such representations should be remains an open one in AM (see Arvaniti, Ladd, and Mennen 2006b; Arvaniti 2016; see also sections 1.5 and 1.6 for a brief discussion).

1.2.4 Scaling and Downstep

Phonologically, AM models rely on two levels, L and H. The use of just two phonological levels that show phonetic variation in scaling avoided issues that had arisen with previous models, such as those of the American structuralists, which relied on several tonal levels (e.g., Pike 1945; Trager and Smith 1951; for an account of this debate, see Arvaniti 2011). Some issues with the AM approach do remain, however.

One of the most problematic aspects is the presence of *downstep*, the lower-than-expected scaling of H tones. In Pierrehumbert (1980) and Beckman and Pierrehumbert (1986), downstep is treated as the outcome of contextual rules, that is, as allophonic. Thus in Pierrehumbert (1980), the second H tone in an HLH sequence is downstepped, while in Beckman and Pierrehumbert (1986), all bitonal pitch accents are said to trigger downstep of a following H tone. Others, however, have argued that downstepped accents differ in meaning from accents that are not downstepped and thus that downstep should be treated as an independent phonological feature to mark the contrast between downstepped and nondownstepped accents, such as !H* and H*, respectively (Ladd 1983, 2008). The issue of whether downstep is a phonological phenomenon or a matter of phonetic scaling remains unresolved for English and has been a matter of debate more generally.

In some AM systems, additional notations are used to indicate differences in scaling. For example, 0% and % have been used to indicate an intermediate level of pitch that is neither high nor low within a given melody and is often said to reflect a return to a default mid-pitch in the absence of tonal specification (e.g., Grabe 1998; Gussenhoven 2005). Still other systems incorporate additional variations in pitch, such as *upstep* (higher-than-expected scaling); for example, Grice, Baumann, and Benzmüller (2005) in their analysis of German intonation notate a boundary tone that has extra-high pitch as ^H%. These practices reflect two main trends: first, the perceived need to notate what may be phonetic detail; second, the awkwardness that mid-level tones pose for analysis, particularly if evidence suggests that they do contrast with H and L tones proper. Such contrast has been proposed for the !H% boundary tone used with Greek *wh*-questions (Arvaniti and Baltazani 2005; Baltazani, Gryllia, and Arvaniti 2020) and the !H- phrase accent used in the routine calling tune of Polish (Arvaniti, Żygis, and Jaskuła 2017). The use of these (and other) diacritics also reflects a more general issue of how to determine what is phonological and what is phonetic in a given intonational system and thus what should or should not be part of the phonological presentation (see also section 1.5).

1.2.5 The Composition of Melodies

Pierrehumbert (1980) includes a grammar for generating English intonational tunes, as shown in figure 1.5. Primitives of the system can combine freely and, in the case of pitch accents, iteratively. With the exception of left edge boundary tones, which are optional, all other elements are required. In other words, a well-formed English tune must include at least one pitch accent followed by a phrase accent and a boundary tone.

The fact that elements combine freely is connected to Pierrehumbert's position that there is no hierarchical structure for tunes (they are a linear string of autosegments);

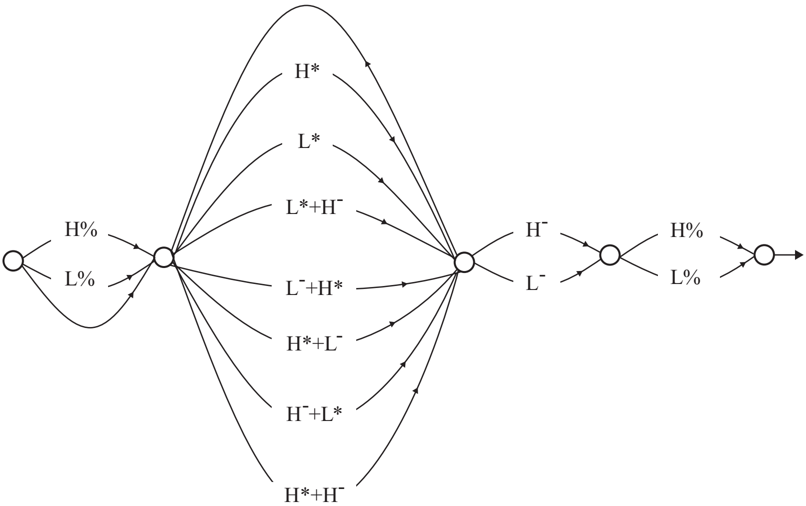


Figure 1.5
The intonational grammar of English. (After Pierrehumbert 1980.)

it follows that there are no qualitative differences between pitch accents and no elements are privileged in any way. Some support for this AM tenet comes from Silverman and Pierrehumbert (1990), who compared the realization of prenuclear and nuclear H* accents in English and concluded that they are phonetically comparable and thus should not be distinguished by virtue of their position in the utterance. Thus, although the term *nucleus* is used in AM to denote the last accent in a melody, AM does not accord the nucleus a special status. This view is juxtaposed to that of the British school, which distinguishes between the "head" and the "nucleus." Simplifying somewhat, the former is the part of the contour before the last pitch accent; the last pitch accent and following edge tones together form the nucleus, considered (as the name implies) the most important and only necessary part of a tune. Despite this difference in approaches, the primacy of the nuclear configuration in AM is indirectly reflected in the grammar shown in figure 1.5 because the required elements for a well-formed minimal tune (in English at least) are a pitch accent, a phrase accent, and a boundary tone, the three elements that together form the British School nucleus.

Not everyone who works within AM shares the view of Pierrehumbert (1980) regarding the nucleus. Gussenhoven (2004, chapter 15; 2005) provides analyses of English and Dutch intonation, respectively, which rely on the notion of nuclear contours as units, a conception close to that of the nucleus adopted by the British School. According to Gussenhoven (2004), the grammar of English intonation relies on three components: nuclear contours, which include the nuclear accents and following boundary tones; the prenuclear pitch accents; and the initial boundary tones or onsets. Ladd (2008, chapter 8) also proposes a model of metrical structure for tunes that divides a tune into three regions: prenuclear, nuclear, and postnuclear. Ladd also argues that this structure is "more abstract than the string of tones" (286) in that tones are not divided by function at this level. In other words, at the level of the tune, tones are simply Ls and Hs without any associations but with a strength relationship between them. The view that the nucleus is somehow privileged is also supported by Calhoun (2010), whose research indicates that nuclear and prenuclear accents may serve distinct functions (see section 1.5).

It is unclear at this stage whether the more elaborate structures that Gussenhoven (2004, 2005) and Ladd (2008) advocated are necessary for all languages, as little relevant work is available. In practice, however, it is clear that certain combinatorial possibilities are extremely rare, while others are very frequent, though there is little systematic investigation of this particular point as well (but see Gussenhoven 2016). Perhaps the most comprehensive study to date is Dainora (2001, 2006) who applied Hidden Markov modeling to the Boston University Radio Speech Corpus (Ostendorf, Price, and Shattuck-Hufnagel 1996) to examine and predict the distribution and combinatorial possibilities of tonal events. Dainora found that some tunes are much more frequent than others; for example, H* L-L% occurred in 33 percent of the corpus, while L*+H L-H% occurred in less than 0.5 percent. She also found a small but significant difference in the transitional probabilities between prenuclear and nuclear accents, which could indicate that the nuclear accents are structurally different from prenuclear accents, as advocated by Gussenhoven (2004, 2005) and Ladd (2008). Dainora's corpus may not be entirely representative of natural speech, as it is based on a relatively small number of radio news announcers reading news stories both in their radio and nonradio styles. Nevertheless, it does indicate that some combinations of tonal elements form standard tunes.

This observation appears to be correct for other languages as well, in that certain meanings appear to be consistently expressed using particular combinations of pitch accents and edge tones. The melodies illustrated in figures 1.1 and 1.2 fall into this category. The fact that some combinations appear to be consolidated into tunes with default functions perhaps should not be surprising: as noted, in AM, tonal entities are morphemes with pragmatic meaning, while the meaning of a tune is created compositionally (Pierrehumbert and Hirschberg 1990; see section 1.4). If so, then it is reasonable to assume that some combinations are preferred and standardized as they reflect frequently used pragmatic meanings. This statement, however, should not be taken to mean that there is a one-to-one relationship between tunes and meanings. I return to this point in section 1.4.

1.2.6 Metrical Structure and Its Relationship with the Autosegmental String

As mentioned in section 1.2.1, the relationship between tonal autosegments and the segmental string (often referred to as *tune-text association*) is mediated by metrical structure. Metrical structure represents both relative strength among constituents and their grouping into phrasal constituents. Both are necessary in the majority of languages investigated so far. At the same time, evidence also shows that while all languages use edge tones to demarcate phrasal constituents, some languages do not make use of pitch accents; this applies, among others, to Korean (Jun 2005a), Chickasaw (Gordon 2005), and Mongolian (Karlsson 2014). In these languages intonation makes reference to phrasing only.

Limited attention has been paid to the link between metrical structure and the autosegmental string (but see, e.g., Gussenhoven 1990 and Cole and Shattuck-Hufnagel 2016 on English; Post 1999 on French; D'Imperio et al. 2005 on several Romance languages; Prieto 2007 on Spanish; Katsika 2016 on Greek, and Gussenhoven 2004 for a review). Perhaps the most developed early model is that of Pierrehumbert and Beckman (1988), which has been implicitly adopted and informally used in many analyses. This model uses *n*-ary branching trees that abide by the Strict Layer Hypothesis (SLH) according to which all constituents of a given level in the hierarchy are exhaustively parsed into constituents of the next level up (Selkirk 1984). However, Pierrehumbert and Beckman (1988, 147ff) also accept limited extrametricality. Crucially, the number of levels in the prosodic hierarchy is not fixed but language-specific. Thus, Pierrehumbert and Beckman posit three main levels of phrasing for Japanese—the AP, the ip, and the IP[2]—but only two levels for English, the ip and the IP, because they found no strong evidence for the AP (Beckman and Pierrehumbert 1986). Finally, the model assumes that it is possible to have headless constituents, that is, constituents that do not include a prominent element. In Pierrehumbert and Beckman's analysis, this applies to Japanese APs that do not include a word with a lexical pitch accent. As noted, the same understanding applies by and large to several other languages that allow headless constituents, such as Korean (Jun 2005a), Chickasaw (Gordon 2005), Mongolian (Karlsson 2014), Tamil (Keane 2014), Bengali (Khan 2014), and West Greenlandic (Arnhold 2014).

A more radical and explicit deviation from the SLH is that advocated by Ladd (2008, chapter 8 and references therein), who argues in favor of compound prosodic constituents. This model allows for limited recursiveness in that constituents of one level in the prosodic hierarchy can form compound constituents of the same level, rather than grouping with constituents of the immediately higher level; for example, two ips can form a compound ip. This is illustrated in (3) with materials from Ladd; the bolded brackets mark the compound domains. Ladd provides several arguments in favor of this alternative, including evidence from the organization of tunes, pitch-accent scaling, and declination reset.

(3) (a) [[[Warren is a stronger campaigner,] [and Ryan has more popular policies,]] [but Allen has a lot more money.]]
(b) [[Warren is a stronger campaigner,] [[but Ryan has more popular policies,] [and Allen has a lot more money.]]]

There are two additional differences between the models discussed and models of prosodic phonology such as Selkirk (1984) and Nespor and Vogel (1986). First, models used in AM fully espouse the idea that prosodic structure above the word is affected by factors such as utterance length, speaking rate, and style. As Beckman and Pierrehumbert (1986, 288) put it, "phrasing in English is highly facultative." In this view, then, phrasing is best determined once a sentence has been uttered and may involve unexpected choices (cf. Arvaniti 2016). Second, the criteria about prosodic phrasing are primarily determined by the tune (though ample evidence is now available from sandhi and other segmental phenomena, as well as durational patterns and articulatory strengthening; see Dilley, Shattuck-Hufnagel, and Ostendorf 1996; Fougeron and Keating 1997; Byrd and Saltzman 1998, 2003; Cambier-Langeveld and Turk 1999; Turk and Shattuck-Hufnagel 2000, 2007; Byrd 2000; Byrd et al. 2000; Cho 2005 on English; Cambier-Langeveld and Turk 1999 on Dutch; Fougeron 2001 on French; Nakai et al. 2009 on Finnish; Themistocleous 2014 on Cypriot Greek). In contrast, phonological models such as Selkirk (1984) or Nespor and Vogel (1986) posit that prosodic structure is determined by reference to syntactic structure. For example, Nespor and Vogel (168) define the domain of the phonological phrase φ (roughly the equivalent of the ip) as consisting of "a *C* [Clitic Group] which contains a lexical head (X) and all *C*s on its nonrecursive side up to the *C* that contains another head outside of the maximal projection of X." These models also include a number of constituents such as the Clitic Group (Nespor and Vogel 1986) based primarily on sandhi and stress patterns. AM is largely agnostic as to the presence of these additional levels of structure, which have rarely been motivated on the basis of intonation.

The issue of metrical strength has received less attention than phrasing within AM. Generally, it is assumed that metrical structure is determined independently of intonation, a position Ladd (2008) dubbed the "stress-first approach" to distinguish it from views advocating an accent-first approach (e.g., Selkirk 1984, section 5.3). The difference has to do with whether metrical structure reflects prominence patterns that drive accentuation (stress-first) or whether accentual patterns are independently determined and drive changes in metrical structure instead (accent-first). As Ladd argues, the fact that it is possible for words to be focused without being accented and for words to be accented without being in focus supports the stress-first approach.

Both options have been supported by empirical evidence. Shattuck-Hufnagel, Ostendorf, and Ross (1994) analyzed instances of the rhythm rule of English, which is often presented as a change in metrical prominence; for example, "Chinese" in "I speak Chinese" is said to be stressed on its ultima, but in "My Chinese teacher was late," it is said to be stressed on this first syllable (Liberman and Prince 1977). Shattuck-Hufnagel, Ostendorf, and Ross (1994) showed that the rhythm rule results not in a change in metrical prominence as such but, rather, in a shift in the position of the pitch accent in words like "Chinese"; the accentual shift does not affect the realization of stress (i.e., vowel quality, duration, or amplitude), supporting the view that metrical structures are independent of and can influence accentuation (see also Gussenhoven 1991; Vogel, Bunnell, and Hoskins 1995; Grabe and Warren 1995; Shattuck-Hufnagel 1995). Similar evidence is provided by Beaver et al. (2007) with respect to "second-occurrence focus" (SOF), the focus (F) on "vegetables" in utterances like "If even [Paul]$_F$ knew that Mary only eats [vegetables]$_{SOF}$, then he should have suggested a different restaurant." Beaver et al. (2007) showed that words with SOF are phonetically prominent, having longer duration and greater energy than nonfoci, but they are not accented.

Furthermore, Calhoun (2010) argues that accentuation cannot always be interpreted as evidence of focus. As Calhoun shows, prenuclear accents are present but often are required for metrical purposes rather than to encode focus and information structure. Calhoun's conclusions about English are strongly supported by similar patterns observed in Spanish (e.g., Prieto, van Santen, and Hirschberg 1995), Greek (Arvaniti and Baltazani 2005), Egyptian Arabic (Chahal and Hellmuth 2014), and Romani (Arvaniti and Adamou 2011), languages in which all content words are typically accented, a pattern that is more plausibly attributed to metrical structure than to information structure requirements.

Metrical requirements can affect accentuation in additional ways. For instance, German, Pierrehumbert, and Kaufmann (2006) have shown that in English, metrical considerations may result in an accent being realized not on the word that represents new information but on a word that is more "accentable." For example, in an exchange like that in (4), the new information in B's utterance is the word "in," which in German, Pierrehumbert, and Kaufmann was accented only 32 percent of the time; in 64 percent of the data, speakers used an accent on "game" instead.

(4) A: Are the children playing their game?
B: Paul took down the tent that they play their game in.

Similarly, Swerts, Krahmer, and Avesani (2002) found that in Italian, words that represent given information remain resolutely accented. Data such as these reinforce the point that accentuation involves language-specific rules rather than universals of what is prominent, new in discourse, informative, or worth highlighting. This conclusion tallies with the focus-to-accent view generally accepted in AM. Informally, this is the position that although accentuation is used to encode focus, the relationship between focus and accentuation is mediated by language-specific patterns and metrical requirements rather than blindly obeying information structure (for a discussion, see Ladd 2008, chapter 6). This view is inherently more compatible with the stress-first rather than the accent-first approach.

1.3 Phonetic Implementation

In terms of phonetic implementation, it is fair to say that in the AM literature, the focus has been on tonal target alignment and scaling. Nevertheless, as mentioned in section 1.1, AM recognizes that intonation does not simply result in the "layering" of an F0 movement on top of particular syllables, but intonation may have additional consequences on articulation, duration, the spectral quality of segments, and the application of sandhi (Beckman and Edwards 1994 and Campbell and Beckman 1997 on English; Baltazani 2006 on Greek; Niebuhr 2008, 2012 on German; Arvaniti, Żygis, and Jaskuła 2017 on Polish). A full discussion of these effects is beyond the scope of this section, which focuses on alignment (section 1.3.1), scaling (section 1.3.2), interpolation (section 1.3.3), and underspecification (section 1.3.4).

1.3.1 Alignment

The idea of tones as targets differentiates AM from models whose primitives are movements (sometimes referred to as *configurations*), such as rises and falls, such as the Institute for Perception Research (IPO) model ('t Hart, Collier, and Cohen 1990) and models that define their phonetic primitives variably as levels or movements, such as PENTA (Xu 2005; Xu et al. 2015). The idea of tonal targets as levels was put to the test in Arvaniti, Ladd, and Mennen (1998), who found that the L and H targets of Greek prenuclear accents each have their own alignment properties. A consequence of this mode of alignment is that the rise defined by the L and H targets has no invariable properties (such as duration or slope). This finding was used by Arvaniti, Ladd, and Mennen (1998) to argue in favor of levels as primitives. Empirical evidence from perception supports this view, showing that listeners perceptually equate pitch movements with level tones (Dilley and Brown 2007).

The strict alignment observed by Arvaniti, Ladd, and Mennen (1998) gave rise to the notion of *segmental anchoring*—the idea that tonal targets anchor to particular segments in phonetic realization. Segmental anchoring was explored in subsequent work by Ladd and colleagues (e.g., Ladd, Mennen, and Schepman 2000; Ladd and Schepman 2003; Ladd et al. 2009). The idea of segmental anchoring also spurred a great deal of research in a variety of other languages (among many, D'Imperio 2001 for Neapolitan Italian; Myers 2003 for Kinyarwanda; Elordieta and Calleja 2005 for Basque Spanish; Arvaniti and Garding 2007 for American English; Dalton and Ní Chasaide 2007 for Irish; Gordon 2008 for Chickasaw; Prieto 2009 for Catalan). This body of research suggests that the idea of segmental anchoring is largely correct, though looser alignment than that reported by Arvaniti, Ladd, and Mennen (1998) is also found (e.g., Welby and Loevenbruck 2006 on French; Smiljanić 2006 on Serbian and Croatian; Arvaniti and Ladd 2009 on Greek). In addition, research has shown that alignment can be dialect-specific (e.g., Grabe et al. 2000 and Fletcher, Grabe, and Warren 2005 on varieties of English; Atterer and Ladd 2004 on varieties of German). Within a linguistic variety, alignment has been shown to be influenced by several factors, including speaking rate, vowel length and weight, and syllable structure (on speaking rate, see Fougeron and Jun 1998 on French; on vowel length and weight, see Schepman, Lickley, and Ladd 2006 on Dutch, and Grice, Roettger, and Ridouane 2015 on Tashlhiyt Berber; on syllable structure, see Prieto and Torreira 2007 on Spanish; Baltazani and Kainada 2015 on Northern varieties of Greek; Petrone and D'Imperio 2015 on Neapolitan Italian; and Grice, Roettger, and Ridouane 2015 on Tashlhiyt Berber). More recently, this increasingly large body of research has also shown that differences in alignment are relevant in encoding and processing phonological contrasts even in tone languages (e.g., Petrone and D'Imperio 2015 on processing in Italian; Remijsen 2013 and Remijsen and Ayoker 2014 on phonological contrasts in Dinka and Shilluk, respectively).

One point that has become very clear thanks to this research is that the traditional autosegmental idea that phonological association necessarily entails phonetic co-occurrence between tone and TBU does not always hold. This applies particularly to pitch peaks. Indeed, one of the most consistent findings in the literature is that of peak delay. This was first documented by Silverman and Pierrehumbert (1990), who examined the phonetic realization of prenuclear H* accents in American English. Empirical evidence for peak delay has since been reported for many typologically unrelated languages with widely different prosodic systems (among many, Prieto, van Santen, and Hirschberg 1995 on South American Spanish; Myers 2003 on Kinyarwanda; Prieto 2005 on Catalan; Dalton and Ní Chasaide 2007 on Irish; Gordon 2008 on Chickasaw). In some instances, other types of alignment have been reported. Perhaps one of the rarest is that reported by Gussenhoven (2000) for Roermond Dutch. Gussenhoven showed that in Roermond Dutch, the realization of a lexical pitch accent on the syllable it is associated with may result in boundary tones being realized nonperipherally, that is, *before* the lexical pitch accent. This pattern, which clearly shows that the linear ordering of tones may not always hold, is analyzed by Gussenhoven (2000) in Optimality Theory terms. It may also be seen as an instance of secondary association the aim of which is to avoid tonal crowding (for a discussion see Gussenhoven 2004, chapter 7).

A particular point of debate has been whether alignment is absolute or proportional to the duration of the anchoring landmark. Evidence for both has been reported. Silverman and Pierrehumbert (1990) advocated in favor of proportional alignment, providing evidence from H* peaks in English with respect to the duration of the accented syllable. Arvaniti, Ladd, and Mennen (1998), on the other hand, argued for absolute alignment of L and H tones in Greek with respect to the accented syllable onset and postaccentual vowel onset, respectively. Schepman, Lickley, and Ladd. (2006) tested various ways of measuring alignment in Dutch and concluded that measuring absolute alignment from a nearby landmark is the most robust. In practice, however, proportional measures can be very useful in comparing alignment across different events: for example, Arvaniti, Ladd, and Mennen (2006b) use proportional alignment to express differences in pitch peak alignment between the Greek L+H* accent and the peak of the H- phrase accent of Greek polar questions. Furthermore, the variety of ways in which alignment is manifested cross-linguistically indicates that the method of alignment of any particular tonal event is best determined empirically by means of controlled materials designed to test predictions based on different alignment modes.

1.3.2 Scaling

Possibly less attention has been given to the scaling of targets compared to alignment. This is partly due to the fact that alignment has been seen as the major diagnostic of tonal composition; for example, the stability of alignment of specific turning points has been used to argue that such points are actual tonal targets and thus reflexes of phonological tones (e.g., Ladd and Schepmann 2003). An additional complication has been the fact that scaling, much more than alignment, reflects not only systematic linguistic differences but also paralinguistic effects; for example, excitement and boredom may result in diametrically different changes in the scaling of targets. A discussion of paralinguistic effects on scaling is beyond the scope of this chapter, which focuses on linguistic differences expected to surface systematically and independently of paralinguistic effects (on paralinguistic uses of pitch, see Gussenhoven 2004, chapter 3; Ladd 2008, section 1.4).

Since Ladd (1996), a distinction is made in AM between *pitch span*, which refers to the extent of the range of frequencies used by a speaker, and *pitch level*, which refers to whether these frequencies are overall high or low; together, level and span reflect two aspects of pitch range. For instance, two speakers may have the same pitch span of 200 hertz but one may use a low level (e.g., 125–325 hertz) and the other a higher level (e.g., 200–400 hertz). It is fair to assume that the pitch level and span of a speaker may change for paralinguistic reasons. Furthermore, language- and gender-specific differences in level and span may also affect tonal scaling (e.g., Daly and Warren 2001; Graham 2014).

With respect to the linguistic aspects of tonal scaling, three main influences are generally agreed on: declination, tonal context, and tonal identity. Following Pierrehumbert (1980), it is generally understood that the scaling of tones can be modeled with reference to a declining baseline that is invariant for each speaker (at a given time). The baseline is defined by its slope and a minimum value that is assumed to represent the bottom of the speaker's range, a value that tends to be very stable for each speaker (Maeda 1976; Menn and Boyce 1982; Pierrehumbert and Beckman 1988). L and H tones (apart from terminal L%s) are scaled above the baseline and with reference to it. For example, Pierrehumbert (1980, 136) notes that the difference between the vocative chant, H*+L- H- L%, and a straightforward declarative, H* L- L%, is that the L% in the former remains above the baseline (because it is upstepped by the preceding H-), while that in the latter does not. The effect of declination is a systematic lowering of targets, though declination can be suspended (e.g., in questions) and is reset across phrasal boundaries (Ladd 1988; see also Truckenbrodt 2002). Listeners do anticipate declination effects and adjust their processing of tonal targets accordingly (e.g., Yuen 2007).

Declination plays a small role in AM as compared to other intonation models, such as IPO ('t Hart, Collier, and Cohen 1990). Recent evidence indicates that this is largely correct in that the relationship between physiology and declination rate is weak (Fuchs et al. 2015). In AM, scaling takes place on the fly with every tone's scaling being calculated as a fraction of the scaling of the preceding tone (Liberman and Pierrehumbert 1984). The exception to this is what Liberman and Pierrehumbert have called "final lowering," the fact that the final peak in a series is scaled lower than what a linear relation between successive peaks would predict. This phenomenon has since been reported in several languages with very different prosodic systems, including Japanese (Pierrehumbert and Beckman 1988), Dutch (Gussenhoven and Rietveld 1988), Yoruba (Connell and Ladd 1990; Laniran and Clements 2003), Kipare (Herman 1996), Spanish (Prieto, Shih, and Nibert 1996), and Greek (Arvaniti and Godjevac 2003). Dialectal variation on the extent and scope of final lowering has also been reported (Arvaniti 2007 on American and British English). Herman, Beckman, and Honda (1996) and Herman (2000) further suggest that final lowering is one of several manifestations of overall "vocal effort" decrease toward the end of an utterance. Less vocal effort results in a drop in subglottal pressure and gesture stiffness, including a relaxation of the muscles of the larynx. These changes can lead to segmental lengthening, a drop in root-mean-square amplitude, voice quality changes, and final lowering, and can be interpreted as indicators of discourse-finality.

With respect to tonal identity, several authors have noted that the scaling of L and H tones is differently affected by changes in pitch span: L tones tend to get lower when pitch span expands, while H tones get higher (e.g., Pierrehumbert and Beckman 1988; Gussenhoven and Rietveld 2000). In addition, consistent differences in the scaling of comparable tonal events have also been reported, suggesting that this is an additional factor to be taken into account during modeling. For instance, Arvaniti and Baltazani (2005) note that in Greek, the peak of H*+L is scaled higher than that of H*. Arvaniti and Garding (2007) report consistently higher scaling of the H tone in American English L*+H as compared to L+H*. Similarly, Vanrell (2007) found that string identical polar and *wh*-questions in Catalan can be disambiguated thanks to the higher scaling of %H in the former as compared to H* in the latter. Similar results are reported by Borràs-Comes, Vanrell, and Prieto (2014) with respect to Catalan L+H*, which can convey different pragmatic meanings depending on peak scaling.

Finally, scaling is affected by tonal context. As mentioned in section 1.2.4, tonal context was the explanation for downstep in Pierrehumbert (1980) and Beckman and Pierrehumbert (1986). Additional interactions between tones have also been reported. Thus, in Pierrehumbert's original analysis, the vocative chant, which ends in sustained level pitch in the middle of a speaker's range, was represented as H*+L- H- L%. In this analysis, the trailing tone of the pitch accent is not phonetically realized but serves to downstep the following H- phrase accent; the L% is in turn upstepped by virtue of being preceded by an H- phrase accent. This type of analysis has not been adopted for other languages, though the representation of sustained phrase-final pitch remains problematic (for discussions see Ladd 2008, chapter 3; Arvaniti 2016; Prieto and Hualde 2016). Interactions such as upstep and downstep as a result of tonal context are certainly not to be taken for granted cross-linguistically; for example, Arvaniti and Baltazani (2005) show that in Greek, neither the presence of bitonal accents nor combinations of L and H edge tones result in downstepping or upstepping of tones. Data such as these are relevant for the treatment of downstep as they indicate that perhaps some languages do require a phonological distinction in scaling that involves three levels rather than just two.

1.3.3 Interpolation

Interpolation has not been investigated as extensively as alignment and scaling have been. It is generally assumed that interpolation is linear between tones, though it is clear that evidence to the contrary is available. The first such piece of evidence is the sagging interpolation briefly discussed in section 1.2.1 (Pierrehumbert 1980, 1981). Sagging interpolation has always been seen as something of an anomaly, leading Ladd and Schepman (2003) to suggest that it is more plausibly analyzed as the reflex of an L tone.

Nonlinear interpolation is arguably evident in the realization of particular tonal events as well. For instance, L* + H and L + H* in English differ in terms of shape, the former being concave and the latter convex, a difference that is neither captured by their autosegmental representations nor anticipated by linear interpolation between the L and H tones. To account for this difference, Barnes et al. (2012, 2013) proposed a new measurement, tonal center of gravity (TCoG), the formula for which is shown in (5).

(5) $$T_{\text{cog}} = \frac{\sum_i f0_i \times t_i}{\sum_i f0_i}$$

The aim of TCoG is to quantify the overall “tilt” or “mass” of a tonal event; for example, L* + H would have a later TCoG than that of L + H*. As Barnes et al. show, TCoG is also relevant in perception, providing, as it were, a perceptual reference location for the event. The existence of the difference quantified by TCoG can be interpreted in two different ways. One possibility is to accept that tonal targets are individual points, as is traditionally maintained in AM, and that therefore interpolation between them can take different shapes, as indicated by the English L* + H and L + H*. In that sense, TCoG captures this type of phonetic difference. Alternatively, one may consider that some tones have duration (see Cole 2000). This is a straightforward way of explaining the concave shape of L* + H, that is to say that it is the result of the longer duration of its L* tone.

The fact that tones may have duration, rather than being turning points, is somewhat problematic for the basic target-and-interpolation view. This applies not only to the L tones just discussed but to H tones as well. H tones are typically said to be realized as peaks, but evidence suggests that plateau realizations are also possible, further arguing for tonal duration. Though in some languages plateaus appear to be used interchangeably with peaks (e.g., Arvaniti 2016 on Romani) in others the two are distinct: the choice of a peak or a plateau affects interpretation of the tune (e.g., D’Imperio 2000; D’Imperio, Terken, and Piterman 2000 on Neapolitan Italian), the scaling of the tones involved (e.g., Knight and Nolan 2006; Knight 2008 on British English), or both (Barnes et al. 2013 on American English). Data such as these indicate that a phonetic model involving only targets as turning points and linear interpolation between them may be too simple to fully model all phonetic detail pertaining to F0 curves or account for its processing by native listeners (but see Prom-On et al. 2016, who found no evidence that using simple AM-based target and interpolation modeling affects listener judgments).

1.3.4 Underspecification

The fact that the phonetic implementation of an AM phonological representation relies solely on the realization of its phonological elements means that melodies are not only phonologically underspecified, but they also show sparse specification at the phonetic level. In other words, in many melodies, such as those in the right column of figure 1.1, most syllables are not assigned a specific F0 value during production. On the other hand, it is also possible for an utterance to involve more tones than TBUs; this is illustrated in the examples in the left column of figure 1.1. Such instances are referred to in AM as *tonal crowding*. This conceptualization of the phonetic realization of melodies differs radically from that of full specification models, such as PENTA (Xu 2005; Xu et al. 2015), which encounter problems with both sparse and dense specification. These are briefly discussed here (for a full discussion and data showing the need for sparse specification see Arvaniti and Ladd 2009, 2015 on Greek, and Arvaniti, Żygis, and Jaskuła 2017 on Polish).

Underspecification was first illustrated with Japanese data by Pierrehumbert and Beckman (1988, 13ff.). They showed that the F0 contours of APs without an accented word could be successfully modeled by positing only one H target, associated with the AP's second mora, and one L target at the beginning of the following AP. The F0 slope from the H to the L target depended on the number of moras between the two. This change in the F0 slope is difficult to model if every mora is specified for F0. In a system such as PENTA, it would be possible to stipulate that the tonal specifications of moras between the H and L targets are of low strength and thus weakly implemented because of the greater strength of the H and L targets at the two ends. This, however, would be an ad hoc solution that is particularly problematic when modeling languages such as Japanese, Korean, and many others in which, as noted in section 1.2.5, headless constituents are possible. In such languages, solutions based on relative tonal strength are not linguistically motivated, rendering any analysis along these lines both arbitrary and circular.

Research in AM has examined not only underspecification but also its flip side, tonal crowding. This body of research shows that tonal crowding is phonetically resolved in a number of ways: (i) *truncation*, the elision of parts of the contour (Bruce 1977; Grice 1995; Arvaniti 1998; Grabe 1998; Grabe et al. 2000; Arvaniti and Ladd 2009); (ii) *undershoot*, the realization of all tones without them reaching their targets (Bruce 1977; Arvaniti, Ladd, and Mennen 1998, 2000, 2006a, 2006b; Prieto 2005; Arvaniti and Ladd 2009); and (iii) temporal realignment of tones (Silverman and Pierrehumbert 1990; Arvaniti and Ladd 2009).[3] Undershoot and temporal realignment often work synergistically, giving rise to compression. Attempts have been made within AM to pin different resolutions of tonal crowding to specific languages (Grabe 1998). Empirical evidence, however, indicates that the mechanism used is specific to elements in a tune rather than to a language as a whole (see Arvaniti and Ladd 2009 on Greek; Arvaniti 2016 on Romani; Rathcke 2017 on German and Russian; see Ladd 2008 for a review). Because of this property of tunes, Arvaniti, Żygis, and Jaskuła (2017) argue that the behavior of putative tonal targets under tonal crowding can serve as a diagnostic of phonological status: elements that can be eliminated under tonal crowding need not be part of the phonological representation.

Dealing with tonal crowding is even more challenging than underspecification for models like PENTA. While in instances of sparse tonal specification, it is possible to assign, albeit arbitrarily, different tonal strengths to particular syllables, this is not possible when the same melody is realized on one syllable only. The problem is both theoretical and practical. From a theoretical perspective arguing, for example, that “Me?!” in figure 1.1 has a rise-fall-rise pitch target misses a significant generalization, namely that the contours of “Me?!” and “A driving instructor?!” are instances of the same melody, in that they do not differ in meaning and function. This is a problem for a model such as PENTA in which tonal specification and communicative functions go hand in hand, as the same communicative function must be arbitrarily linked to a seemingly arbitrary assortment of contours. Pitch contours like those of “Me?!” are problematic from the perspective of modeling as well: with only one syllable in place, it is unclear how the relative importance of each part of the melody is to be modeled. Yet the fact that different aspects of a contour are prioritized based on the melody is undisputable. This is illustrated in figure 1.6 with data from the urgent call melody of Polish (Arvaniti, Żygis, and Jaskuła 2017). Words with different degrees of tonal crowing are shown: “Piotr,” which exemplifies the most extreme tonal crowding; “Jaś”; and “Marianna,” which does not involve tonal crowding and is added here for comparison. As can be seen, although both the onset and coda are voiceless in “Piotr,” the rise is truncated, but the fall is not; it is only slightly undershot by comparison to “Marianna.” This difference cannot be accounted for by relative tonal strength, because in PENTA, both the rise and fall are tonal specifications connected to the same syllable. In AM, however, the analysis of the rise as the reflex of an H* pitch accent, which is independently motivated, allows for this difference. The rise is not part of the tonal specification of the tune but simply interpolation between a default starting

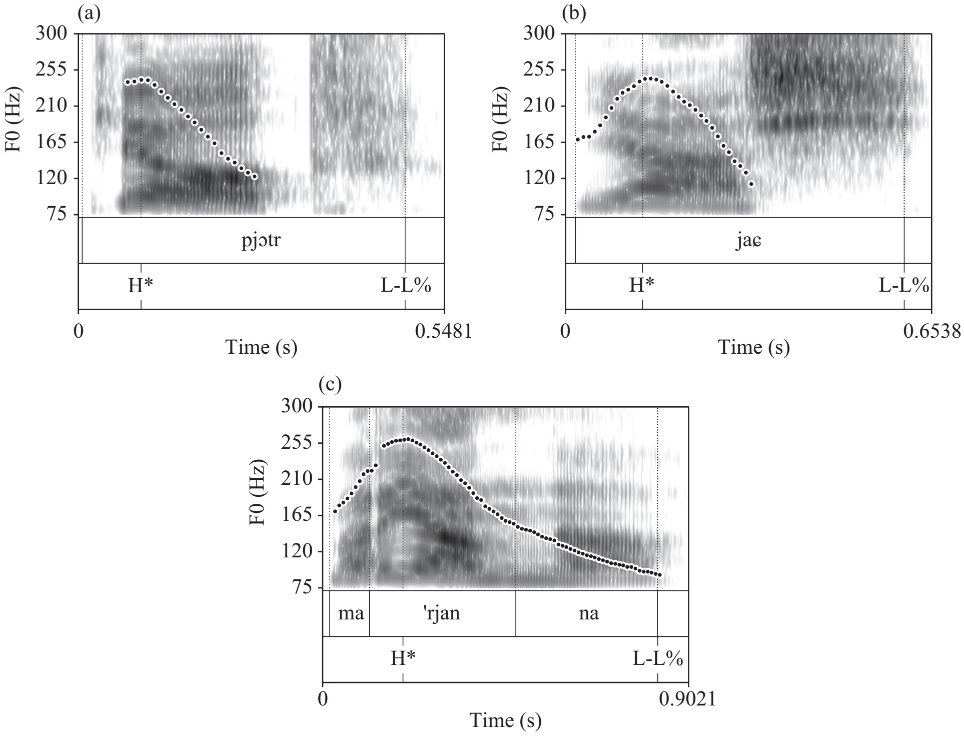


Figure 1.6

Spectrograms and F0 contours of the urgent call of Polish, illustrating the effects of different degrees of tonal crowding on the realization of the tune. (Adapted from Arvaniti, Żygis, and Jaskuła 2017.)

level, attested in other Polish contours as well, and the target (the peak); thus, the rise can be truncated under extreme tonal crowding like that of "Piotr," but the L% boundary tone at the right edge of the utterances cannot (for additional discussion and evidence, see Arvaniti, Żygis, and Jaskuła 2017).

1.4 Intonation and Meaning

Intonational meaning has been the Achilles heel of many models of intonation. One of the major difficulties that most models have encountered is the fact that the meaning of a tune appears to change dramatically depending on context. This can be illustrated with an example from Bolinger (1964): "Because I wanted to" can be used with rising intonation as a petulant response to "Why did you do it?," but the same tune can also be used for a straight forward question, as in "Was it yesterday?"[4] The difficulty of handling such seemingly arbitrary changes in meaning and modality is evident in the elusiveness of phenomena such as "ironic prosody" (e.g., Bryant and Fox Tree 2005) or the treatment of intonational meaning in the British school, which essentially amounted to attaching to each tune a long list of attitudinal meanings, often quite contradictory from a pragmatic perspective; for example, one of the many meanings of the "low drop" of O'Connor and Arnold (1973, 49) is "to give weight to expressions of both approval and disapproval, of both enthusiasm and impatience." Other systems either do not deal with meaning (e.g., IPO; 't Hart, Collier, and Cohen 1990), or limit the exploration of intonational meaning to simple "communicative functions" such as statement versus question (e.g., Xu 2005, Xu et al. 2015). Finally, intonational meaning has been studied quite extensively in research on information structure, where, however, intonation has often been treated largely as a vehicle for focus with little recognition of its contribution to pragmatics (among many, Selkirk 1984; Beaver et al. 2007; Féry and Kügler 2008; Buring 2013).

An additional obstacle to understanding intonational meaning has been the widely held view that there is a natural relationship between tunes and meaning (as advocated, e.g., by Ohala 1983 and partly espoused by Gussenhoven 2004, chapter 5). This view has often blurred the boundaries between linguistic and paralinguistic uses of F0 (rather than intonation per se), despite evidence to the contrary. For instance, Chen, Gussenhoven, and Rietveld (2004) have shown that even the perception of paralinguistic information derived from F0 cues is language-specific, as is the encoding of very common intonation functions such as continuation (e.g., Chen 2007). Similarly, research on uptalk has shown that not only does uptalk differ subtly in its realization across varieties of English, it also serves different pragmatic functions (e.g., Fletcher, Grabe, and Warren 2005; Ritchart and Arvaniti 2014; see Warren 2015 for a review). Even one of the most frequently cited pieces of evidence in support of a natural relationship between F0 and modality, namely that questions end in rises, is only weakly supported by cross-linguistic evidence; questions ending in falls or in low F0 are attested in (at least) several African languages as shown by Rialland's (2007) survey, as well as in Greek, Hungarian, Romanian, and Cypriot (Grice, Ladd, and Arvaniti 2000), Romani (Arvaniti 2016), and Tashlhiyt Berber (Grice, Roettger, and Ridouane 2015).

Many of these issues are dealt with successfully in the model of Pierrehumbert and Hirschberg (1990), perhaps the best-known treatment of meaning within AM. Though their analysis is specific to English, its principles are applicable to intonational systems at large. According to this model, each intonational element is seen as a morpheme that encodes pragmatic meaning. The meaning conveyed by an entire melody is therefore compositional, not holistic as in other models, and depends on contributions from all elements of the melody. One advantage of this model is that it can account for similarities in meaning conveyed by identical components of different melodies, such as the use of the same pitch accent with different edge tones. At the same time, by not relying on some natural connection between tones and meaning, it is more easily amenable to an understanding of cross-linguistic and cross-dialectal variation in the relationship between intonational meaning and form.

Pierrehumbert and Hirschberg treat intonation as a means speakers use to specify a particular relationship between the propositional content of their utterance and the mutual beliefs of the discourse participants. These are beliefs that participants come to share as a result of the conversational interaction. Thus, the use of a particular tune can indicate what a speaker may think is or is not part of mutual beliefs (though Pierrehumbert and Hirschberg also make use of one-sided mutual beliefs, a particular speaker's beliefs about what is mutually believed by all participants). In this framework, intonation does not have an attitudinal function (as in the British School, for instance), nor is it assumed to have a one-to-one relationship with meaning, such that a specific communicative function is expressed using a particular tune (as in PENTA). As Pierrehumbert and Hirschberg (284) note, "Attitude is better understood as derived from tune meaning interpreted in context than as presenting that meaning itself." In other words, the meaning of a tune is not to be directly interpreted; for example, H* L-L%, often used with declaratives, is not to be interpreted as "the speaker believes *x*"; rather, the speaker's belief in *x* will be inferred from all the components of the tune and the context in which they are used. For example, "That's fantastic!" uttered with the H* L-L% tune after the speaker receives bad news will not be interpreted as a declaration that the speaker believes the news is fantastic, but as a sarcastic utterance declaring the exact opposite. This approach allows AM to account for the different interpretations of the same tune in different contexts (cf. Baltazani, Gryllia, and Arvaniti 2020). This is because meaning is compositional, sufficiently abstract, and based on the idea that in a conversation, speaker and hearer contribute and alter mutual beliefs.

The principles of Pierrehumbert and Hirschberg can be illustrated with an analysis of Bolinger's (1964) examples "Because I wanted to" and "Was it yesterday?" In AM terms, the melody of both is L* H-H%. According to Pierrehumbert and Hirschberg, items accented with an L* pitch accent are not part of what the speaker is predicating about the utterance and thus they are not to be added to the interlocutors' system of mutual beliefs. The use of H-H% conveys "forward reference" (306); informally, we could say that by using H-H%, the speaker throws the ball back to the listener. Based on the preceding discussion, it is easy to see why L* H-H% would be used with the question: the speaker highlights *yesterday* using L* but at the same time invites the hearer to make a predication, she is not making one herself; H-H% indicates the expectation of an answer. In short, in this instance, the speaker offers no predication but invites the listener to do so. Similarly, L* in "Because I wanted to" indicates that the speaker does not really commit herself to the answer, while H-H% invites a response. In short, by using L* H-H% with "Because I wanted to," the speaker conveys that she does not wish to commit to a proper explanation of her actions and at the same time challenges the speaker to respond.

Despite the existence of this framework, it is fair to say that many aspects of intonational meaning remain unclear within AM. Pierrehumbert and Hirschberg themselves note that similarities between the meanings of bitonal accents (L*+H and L+H*, and H*+L and H+L*) are evident in English but not easy to account for in their model. As with the analysis of Pierrehumbert and Hirschberg, newer analyses, such as Steedman (2014), have not been tested with extensive and natural corpora. Furthermore, models similar to that of Pierrehumbert and Hirschberg have not been developed for other languages, so the framework has not been extensively tested for cross-linguistic validity (though research in this area is becoming more widespread; e.g., Armstrong and Prieto 2015 on Puerto Rican Spanish; Borràs-Comes, Sichel-Bazin, and Prieto 2015 on Catalan; Baltazani, Gryllia, and Arvaniti 2020 on Greek). Researchers have also found that certain configurations appear to be more frequent than others (e.g., Dainora 2001, 2006). Observations of this sort have given rise to hypotheses that intonational meaning may depend on groups of tones (Gussenhoven 2004; Peters 2018) or abstract tunes (Ladd 2008), rather than being strictly compositional (for reviews and discussion, see Gussenhoven 2004, chapter 7; Ladd 2008, chapter 8). Those who argue that tones form groups assume a hierarchical structure of tunes and assign priority to the meaning of the *nucleus*, the part of the tune that consists of the last pitch accent and following edge tones (Gussenhoven 2004; Ladd 2008). Though these approaches highlight the importance of the nuclear configuration for intonational meaning, more recent evidence indicates that prenuclear accents can also convey pragmatic information (e.g., Petrone and D'Imperio 2011 on Neapolitan Italian; Petrone and Niebuhr 2014 on German; Baltazani et al. 2015 on Greek). As this discussion indicates, phonological assumptions about the structure of tunes interact with and affect assumptions about the nature and structure of meaning. Thus, it is not possible to reach satisfactory conclusions about one without properly scrutinizing the other.

As this brief discussion indicates, much remains to be done before the relationship between intonation and meaning is fully understood. It is clear, however, that crude distinctions such as that between statement and question, or focus and lack thereof, are not sufficient to explain intonational meaning, and that a more sophisticated understanding is required, possibly along the lines of Pierrehumbert and Hirschberg (1990), or with their analysis as a starting point. Finally, a principled distinction between the pragmatic functions of intonation, on the one hand, and the role of accentuation in encoding focus, on the other, together with a recognition that both are part of the functions of intonation, are likely to lead to greater insights into intonational meaning.

1.5 Intonational Typology

Typology is frequently discussed in the AM literature. However, there appears to be no generally accepted view of what this typology would address, particularly with respect to the tonal aspects of AM. Jun (2005b, 2014a) provides two related proposals of a general prosodic typology, which includes both a categorization of word-level prosody and a categorization of postlexical prosodic events. Jun (2005b) groups both by prominence, which she takes to be a category of prosodic typology. For the lexical level, Jun accepts the traditional division among tone languages (e.g., Mandarin), stress-accent languages (e.g., English), and lexical pitch-accent languages (e.g., Japanese). Languages are said to be tonal if "pitches [have] paradigmatic contrast" (440). In stress-accent languages, one syllable in a word is made more prominent by means of duration and/or amplitude, setting up a syntagmatic contrast between syllables. Finally, languages with lexical pitch accent are those in which one syllable in a word has lexical specification for pitch but is not otherwise made prominent relative to other syllables.

An additional category is implicitly accepted by Jun (2005b) to include languages such as Korean, which has neither stress-accent nor tonal specification of any type at the lexical level, and thus does not fit any of the three traditional types. At the postlexical level, Jun proposes a distinction between languages that use prominence culminatively by marking the head of a prosodic constituent and those that use prominence demarcatively, by marking its edges. One can informally equate these two types with accentuation and edge tones, respectively. Jun argues persuasively that in languages like Korean, which are of the latter type, "the function of postlexical pitch accent in

English and other West Germanic languages is performed by prosodic phrasing" (441). The difference between edge and head marking is predictive: in languages like English and other languages that mark prosodic heads using intonation, focus is marked using accentuation with concomitant deaccenting of subsequent items in the utterance; in languages like Korean, the intonational encoding of focus involves instead dephrasing (e.g., Jun 2005a for Korean, Venditti 2005 for Japanese).

Jun (2005b) also includes rhythm as a dimension in her typology, making use of the traditional rhythm classes: syllable-, stress-, and mora-timing (on reasons why this typology is problematic and now largely abandoned, see Arvaniti 2009b, 2012). The final dimension is that of prosodic constituents and their size, a typological difference Jun (2005b) links to rhythm class. Jun (2014a) recognizes some issues with this typology, one of the most significant being that it does not provide a way to directly compare tonal aspects of prosody. To account for this, Jun (2014a) proposes an additional concept for her typology, that of *macro-rhythm*, which is based on changes in F0 and is used together with *micro-rhythm*, which encompasses the traditional rhythm classes. Macro-rhythm aims to capture the observation that some languages, such as English, use accentuation sparingly, while others, such as Greek or Spanish, accent every word (see section 1.2.6). Three criteria are used to define macro-rhythm: the number of possible phrase-medial tonal events (e.g., different types of pitch accents), the form of tonal events involved (roughly, level tones or glissandi), and the frequency of those events. Thus, a language that has a large inventory of rising and falling pitch accents that are used to accent every word would be considered more macro-rhythmic than a language that sparingly used only level tonal events.

There are some issues with Jun's proposals both for the lexical and postlexical levels. First, there is no agreement among researchers with respect to the prosodic types posited for the lexical level. Hyman (2006, 2009) argues against the lexical pitch-accent category, which he considers vague and lacking a prototype, leading him to argue for only two essential types: tone languages and stress-accent languages (with languages like Japanese probably being treated as tonal languages with privative tone). Beckman and Venditti (2011), on the other hand, argue that languages traditionally classified as tonal and those classified as having stress accent share many similarities in terms of prosodic grouping and the tonal expression of focus; at the same time, both groups are quite distinct from lexical pitch-accent languages like Japanese, which pattern with languages without prominence at the lexical level, such as Korean. The analysis of Beckman and Venditti rests on the idea that the function and provenance of tones at the lexical level are best seen as predictors for the function and role that tonal elements play at the postlexical level, rather than as predictors of types in and of themselves.[5] In this respect they agree with Jun (2005b, 2014a) by giving priority to the edge versus head prominence parameter, which is better expressed by their typological grouping.

Second, there are some issues with the idea of macro-rhythm. First, its predictive power and relevance for the organization of prosody (if any) are at present unclear. In addition, there is currently little quantitative evidence for macro-rhythm, so it is also unclear whether the combination of Jun's (2014a) three criteria would result in different types in practice (e.g., if macro-rhythm were to be quantified with data from natural speech corpora). This is of some concern for two reasons. First, some of the criteria may cancel each other out; for example, English does have a variety of pitch accents, but typically the same accent is used phrase-medially in an utterance (Dainora 2006), so variety may apply at the inventory level but may not apply in practice. Second, some of the criteria for macro-rhythm may not be completely independent of each other; for example, if a language uses only level tonal events, the tonal inventory is likely to be small, and this, in turn, can limit variability in F0 contours.

Ladd (2008, chapter 3), on the other hand, has argued that the difficulty in making meaningful typological comparisons is related to the variable representations used among AM analyses of different languages for seemingly similarly phenomena. This argument reveals two main issues that plague typological research: the nature of AM representations and criteria for similarity in intonation. First, there is no consensus about the necessary level of abstraction for a typology of intonation. It would seem plausible for Ladd to consider abstract tunes (see section 1.2.5) as basic for typology. Yet his main example in this discussion is sustained-level phrase-final pitch, which is a rough and ready phonetic description. As the discussion throughout this chapter has indicated, a pitch contour can be the reflex of a variety of tones of both lexical and postlexical provenance; no two languages break up a pitch contour in identical ways, and as a result, a particular pitch configuration, such as sustained-level phrase final pitch, can come about in different ways. In other words, the fact that they may look similar does not necessarily make two contours instantiations of the same phenomenon. This serves as a reminder that as argued by Hyman (2006), any phonological typology must deal not with surface phonetic details but rather with the analytical categories used to make sense of these details in a given linguistic system. Though concrete solutions do not appear at present to be forthcoming, at the very least, it is important in any attempts at typology not to confuse the two (Ladd 2008; Arvaniti 2016). Applying Hyman's principle to intonation would mean that typologically relevant elements are unlikely to be rising versus falling pitch or details about peak alignment or the scaling of tones. This in turn suggests that any categories used for typology cannot be determined a priori for all languages and be based on surface phonetic representations, as argued by Prieto and Hualde (2016). Rather, they are likely to be more abstract concepts, allowing for language-specific instantiations, as argued by typologists such as Haspelmath (2010, 2015). If this assessment is correct, then parameters such as edge versus head prominence are more likely to be relevant for understanding how tones are organized and function in different systems (Jun 2005b; Jun and Fletcher 2014;, Beckman and Venditti 2011).

This debate on the role of typology brings to the fore one major issue in the field: the degree of abstraction required for phonological representations of AM. Many of the issues discussed here reflect precisely this type of disagreement. This includes the role of secondary associations to phonologically represent phonetic alignment, the use of diacritics to represent degrees of scaling or alignment details, and the treatment of downstep. As I have argued elsewhere, some of these issues can be resolved if the phonological nature of intonation is accepted and intonation is treated in a similar fashion to other subsystems of phonology (Arvaniti 2016, 2019). This would require accepting that phonetic form is not the end-all of intonational research, that categories must be posited only if contrastive in a system, and that contrast must involve distinctions in meaning, not just in form, as "a significant difference between two populations of signals does not necessarily indicate the nature of the phonological difference" (Shaw and Kawahara 2018, 512). This line of argument also implies recognition that phonetic realization will be variable (due to contextual factors) and must also be expected to show dialectal and sociolinguistic variability, much as it happens with all elements posited in any phonological system (Arvaniti 2019).

1.6 Applications of AM: ToBI Systems and Transcription

The best-known application of AM is the family of ToBI systems. ToBI (tones and break indices) was a system originally developed for the prosodic annotation of American English corpora for use in speech synthesis (Silverman et al. 1992; see also Beckman, Hirschberg, and Shattuck-Hufnagel 2005; Brugos, Shattuck-Hufnagel, and Veilleux

2006). Since then, several similar systems have been developed for a variety of languages (see, e.g., Jun 2005c, 2014a for relevant surveys). To distinguish the system for American English from the general concept, the term MAE_ToBI is now used for the former (where MAE stands for Mainstream American English; Beckman, Hirschberg, and Shattuck-Hufnagel 2005). As mentioned, ToBI was originally conceived as a tool for research and speech technology; for example, a ToBI annotated corpus can be searched for instances of the same intonational events, such as the H* accent in English, so that a set of its instantiations can be analyzed and generalizations as to its realization reached. Such generalizations are useful not only for speech synthesis but also for phonological analysis and the understanding of variation (for additional uses and extensions, see Jun 2005b).

ToBI representations consist of a sound file, associated spectrogram and pitch track, and several tiers of annotation. The required tiers are tones—a representation of the tonal structure of the pitch contour—and the break index tier, in which the perceived strength of prosodic boundaries is annotated using numbers; for example, in the MAE_ToBI system, [0] represents cohesion between items (such as flapping between words in American English); [1] represents the small degree of juncture expected between most words; [3] and [4] represent ip and IP boundaries, respectively, and [2] is reserved for uncertainty (e.g., for cases when the annotator cannot find tonal cues for the presence of a phrasal boundary but does perceive a high degree of juncture). Additional tiers may include an orthographic tier and a miscellaneous tier for additional information, such as disfluencies. The content of all tiers is adapted to the prosodic features of the system under analysis but also to particular research needs and theoretical positions of the developers. For instance, Greek GRToBI marks sandhi, which is widespread in Greek and thus of some interest (Arvaniti and Baltazani 2005), while Korean K_ToBI includes both a phonological and a phonetic tier (Jun 2005a), as does the IViE (Intonational Variation in English) corpus of British English (Nolan and Post 2014).

ToBI as a concept has often been misunderstood. Some consider ToBI to be one among many transcription systems for intonation, similar to, say, INTSINT (or International Transcription System for Intonation; Hirst, Di Cristo, and Espesser 2000), which is a system of symbolic notation based on a limited number of symbols and agreed-on conventions for their use, the aim of which is to reproduce the course of F0. Due to its relative popularity, some have taken ToBI to be the equivalent of the IPA for intonation, a claim that the developers of ToBI have taken pains to deny (e.g., Beckman, Hirschberg, and Shattuck-Hufnagel 2005; Beckman and Venditti 2011). It is perhaps this belief that ToBI is a transcription system designed originally and best suited for English that gives rise to frequent assertions that ToBI is not right for certain languages. Yet neither the categories nor the tiers of MAE_ToBI have to be adopted by any other ToBI system; the fact that the overall architecture has been used for languages as distinct as Cantonese (Wong, Chan and Beckman 2005), Japanese (Venditti 2005), Bengali (Khan 2014), and West Greenlandic (Arnhold 2014) testifies to this. Furthermore, a ToBI annotation system is meant to rely on a prosodic analysis of the language in question (though it could be used as an informal tool to kick-start such an analysis on the understanding that annotations would have to be revisited once the phonological analysis is complete; Jun and Fletcher 2014; Arvaniti 2016).

The confusion about ToBI arises from the fact that it is not always clear that AM is a phonological model of intonational structure, not a transcription, and that ToBI is a tool. Opinions on this point are divided within AM as well, as is reflected in the development of ToBI itself, which is a hybrid between the very abstract system of Pierrehumbert (1980) and more concrete representations such as the explicit marking of downstep which has replaced some of Pierrehumbert's more abstract solutions. It is also reflected in the debate as to whether a level of phonetic transcription is needed in ToBI systems. Recently Prieto and Hualde (2016) have argued for the formalization of this idea, so that AM "labels" be used for an international prosodic alphabet with agreed conventions on the use of a limited set of symbols with prespecified phonetics. Others have argued that the phonetic representation of intonation is the pitch track (Beckman and Venditti 2011; Arvaniti 2016) and that such phonetic transcriptions as advocated by Prieto and Hualde should at best serve as a tool in the early stages of analysis. Arvaniti (2016) has further argued that using AM as a system of broad phonetic transcription is problematic, as doing so turns one symbolic representation into another, the equivalent of the systematic phonetic level of generative phonology criticized by Ladd (2011) with respect to the representation of segmentals.

What needs to remain clear from this discussion is that AM is fundamentally a phonological system and as such relies on the combined investigation of form and meaning. This is one of the major features that distinguishes it from other systems for the modeling and analysis of intonation. As argued elsewhere (Arvaniti and Ladd 2009), this is what gives AM predictive power and allows it to reach useful generalizations about intonation and its relation to the rest of prosody.

Notes

1. Pitch and F0 are often used as synonyms in the intonation literature. Strictly speaking, *F0* refers to the physical property of the speech signal that is directly related to the rate of vibration of the vocal folds, while *pitch* refers to the percept of F0. F0 is measured in hertz. There are several scales for measuring pitch but no strong consensus on which is best for the investigation of intonation. Some studies use hertz, because the relationship between F0 and pitch is linear up to approximately seven hundred hertz, according to the Mel and Bark scales (Nolan 2003), and F0 rarely exceeds seven hundred hertz (except in the voices of infants and toddlers). The two pitch scales used most frequently are equivalent regular bandwidth (ERB) and semitones. ERB reflects a semilogarithmic relation between pitch and F0 in the frequencies used for intonation and has been shown to accurately capture the percept of pitch (Glasberg and Moore 1990; Hermes and van Gestel 1991). Semitones are a logarithmic transformation of the hertz scale primarily used in Western music and now increasingly in intonation research. Evidence in favor of semitones comes from Nolan (2003), who concluded that semitones correspond somewhat better to intonation differences than the ERB scale does (in that the intonation of participants' imitations of utterances matched somewhat more closely the original intonation when expressed in semitones than ERB). Semitones have the dubious advantage of eliminating differences between male and female speakers; this is convenient for statistical analysis, but it may well hide systematic differences between genders that could come to light under separate analyses. Because such differences are valid (e.g., Daly and Warren 2002), eliminating them from analysis does not seem advisable, at least when addressing some research questions.

2. In Japanese J_ToBI, a model of Japanese intonation based on Pierrehumbert and Beckman (1988), the use of the ip is abandoned in favor of two levels only: the AP and the IP (Venditti 2005).

3. The avoidance of tonal choices that would result in crowding is also discussed by Ladd (2008, chapter 5) with respect to German data, but this strategy has not been experimentally investigated to my knowledge.

4. It was precisely this problem with intonational meaning that prompted Pike (1945, 23–24) to caution against the practice of investigating the meaning of tunes using only utterances with the same grammatical structure (such as pitting statements against questions). Doing so can give the false impression that a tune has a consistent meaning.

5. This may go some way toward explaining the disagreement between Hyman (2009) and Beckman and Venditti (2011), because Hyman places emphasis on lexical phenomena, while Beckman and Venditti focus on the postlexical level.

References

Armstrong, Megan, and Pilar Prieto. 2015. "The Contribution of Context and Contour to Perceived Belief in Polar Questions." *Journal of Pragmatics* 81:77–92.

Arnhold, Anja. 2014. "Prosodic Structure and Focus Realization in West Greenlandic." In Jun 2014b, 216–251.

Arvaniti, Amalia. 1992. "Secondary Stress: Evidence from Modern Greek." In *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, edited by Gerard J. Docherty and D. Robert Ladd, 398–423. Cambridge: Cambridge University Press.

Arvaniti, Amalia. 1998. "Phrase Accents Revisited: Comparative Evidence from Standard and Cypriot Greek." In *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, November 30–December 4, 1998, 2883–2886; ISCA Archive, http://www.isca-speech.org/archive/icslp.

Arvaniti, Amalia. 2003. "Peak Scaling in Greek and the Role of Declination." In *Proceedings of Fifteenth International Conference of Phonetic Sciences,* Barcelona, edited by M. J. Sole, D. Recasens, and J. Romero, 2269–2272; ICPhS Archive. http://www.internationalphoneticassociation.org/icphs/icphs2003.

Arvaniti, Amalia. 2007. "On the Presence of Final Lowering in British and American English." In Gussenhoven and Riad, 317–347.

Arvaniti, Amalia. 2009a. "Greek Intonation and the Phonology of Prosody: Polar Questions Revisited." In *E-proceedings of the 8th International Conference of Greek Linguistics (8ICGL),* edited by G. K Giannakis, M. Baltazani, G. I. Xydopoulos, and T. Tsaggalidis, 558–573. Department of Philology: University of Ioannina. http://www.linguist-uoi.gr/cd_web/arxiki_en.htm.

Arvaniti, Amalia. 2009b. "Rhythm, Timing and the Timing of Rhythm." *Phonetica* 66:46–63.

Arvaniti, Amalia. 2011. "The Representation of Intonation." In *Companion to Phonology*, edited by Marc van Oostendorp, Colin Ewen, Elizabeth Hume, and Keren Rice. Hoboken, NJ: Wiley-Blackwell.

Arvaniti, Amalia. 2012. "The Usefulness of Metrics in the Quantification of Speech Rhythm." *Journal of Phonetics* 40:351–373.

Arvaniti, Amalia. 2016. "Analytical Decisions in Intonation Research and the Role of Representations: Lessons from Romani." *Laboratory Phonology* 7 (1): 1–43.

Arvaniti, Amalia. 2019. "Crosslinguistic Variation, Phonetic Variability, and the Formation of Categories in Intonation." In *Proceedings of International Congress of Phonetic Sciences,* Melbourne, Australia, edited by Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren, 1–6. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Arvaniti, Amalia, and Evangelia Adamou. 2011. "Focus Expression in Romani." In *Proceedings of the 28th West Coast Conference on Formal Linguistics*. Cascadilla Proceedings Project.

Arvaniti, Amalia, and Mary Baltazani. 2005. "Intonational Analysis and Prosodic Annotation of Greek Spoken Corpora." In Jun 2005c, 84–117.

Arvaniti, Amalia, and Gina Garding. 2007. "Dialectal Variation in the Rising Accents of American English." In *Papers in Laboratory Phonology 9*, edited by Jennifer Cole and José Ignacio Hualde, 547–576. Berlin: Mouton de Gruyter.

Arvaniti, Amalia, and Svetlana Godjevac. 2003. "The Origins and Scope of Final Lowering in English and Greek." In *Proceedings of the Fifteenth International Congress of Phonetic Sciences, Barcelona*, edited by M. J. Sole, D. Recasens, and J. Romero, 1077–1080; ICPhS Archive. http://www.internationalphoneticassociation.org/icphs/icphs2003. Arvaniti, Amalia, D. Robert Ladd, and Ineke Mennen. 1998. "Stability of Tonal Alignment: The Case of Greek Prenuclear Accents." *Journal of Phonetics* 26:3–25.

Arvaniti, Amalia, D. Robert Ladd, and Ineke Mennen. 2000. "What Is a Starred Tone? Evidence from Greek." In Broe and Pierrehumbert, 119–131.

Arvaniti, Amalia, D. Robert Ladd, and Ineke Mennen. 2006a. "Phonetic Effects of Focus and 'Tonal Crowding' in Intonation: Evidence from Greek Polar Questions." *Speech Communication* 48:667–696.

Arvaniti, Amalia, D. Robert Ladd, and Ineke Mennen. 2006b. "Tonal Association and Tonal Alignment: Evidence from Greek Polar Questions and Contrastive Statements." *Language and Speech* 49:421–450.

Arvaniti, Amalia, Marzena Żygis, and Marek Jaskuła. 2017. "The Phonetics and Phonology of the Polish Calling Melodies." *Phonetica* 73:342–365.

Atterer, Michaela, and D. Robert Ladd. 2004. "On the Phonetics and Phonology of 'Segmental Anchoring' of F0: Evidence from German." *Journal of Phonetics* 32:177–197.

Baltazani, Mary. 2006. "Focusing, Prosodic Phrasing, and Hiatus Resolution in Greek." In Goldstein, Whalen, and Best, 473–494.

Baltazani, Mary, Stella Gryllia, and Amalia Arvaniti. 2020. "The Intonation and Pragmatics of Greek wh-Questions." *Language and Speech* 63:56–94. Article first published online January 25, 2019. https://doi.org/10.1177/0023830918823236.

Baltazani, Mary, and Evia Kainada. 2015. "Drifting without an Anchor: How Pitch Accents Withstand Vowel Loss." *Language and Speech* 58 (1): 84–115.

Baltazani, Mary, Evia Kainada, Aggelos Lengeris, and Katerina Nikolaidis. 2015. "The Prenuclear Field Matters: Questions and Statements in Standard Modern Greek." In *Proceedings of the Eighteenth International Congress of Phonetic Sciences*.

Barnes, Jonathan, Alejna Brugos, Stefanie Shattuck-Hufnagel, and Nanette Veilleux. 2013. "On the Nature of Perceptual Differences between Accentual Peaks and Plateaux." In *Understanding Prosody: The Role of Context, Function and Communication*, edited by Oliver Niebuhr, 93–118. Berlin: Mouton de Gruyter.

Barnes, Jonathan, Nanette Veilleux, Alejna Brugos, and Stefanie Shattuck-Hufnagel. 2010. "Turning Points, Tonal Targets, and the English L- Phrase Accent." *Language and Cognitive Processes* 25:982–1023.

Barnes, Jonathan, Nanette Veilleux, Alejna Brugos, and Stefanie Shattuck-Hufnagel. 2012. "Tonal Center of Gravity: A Global Approach to Tonal Implementation in a Level-Based Intonational Phonology." *Laboratory Phonology* 3 (2): 337–383.

Beaver, David, Brady Zack Clark, Edward Flemming, T. Florian Jaeger, and Maria Wolters. 2007. "When Semantics Meets Phonetics: Acoustic Studies of Second-Occurrence Focus." *Language* 83:245–276.

Beckman, Mary E., and Janet Edwards. 1994. "Articulatory Evidence for Differentiating Stress Categories." In *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, edited by Patricia A. Keating, 7–33. Cambridge: Cambridge University Press.

Beckman, Mary, Julia Hirschberg, and Stefanie Shattuck-Hufnagel. 2005. "The Original ToBI System and the Evolution of the ToBI Framework." In Jun 2005c, 9–54.

Beckman, Mary E., and Janet Pierrehumbert. 1986. "Intonational Structure in English and Japanese." In *Phonology Yearbook 3*, edited by Colin Ewen and John Anderson, 255–310. Cambridge: Cambridge University Press.

Beckman, Mary E., and Jennifer J. Venditti. 2011. "Intonation." In *The Handbook of Phonological Theory*, edited by John Goldsmith, Jason Riggle, and Alan C. L. Yu, 485–532. Hoboken, NJ: Blackwell.

Bolinger, Dwight. 1964. "Around the Edge of Language: Intonation." *Harvard Educational Review* 34 (2): 282–293. Reprinted in Bolinger, Dwight. 1972. *Intonation*, 19–29. Harmondsworth, UK: Penguin.

Borràs-Comes, Joan, Maria del Mar Vanrell, and Pilar Prieto. 2014. "The Role of Pitch Range in Establishing Intonational Contrasts." *Journal of the International Phonetics Association* 44 (1): 1–20.

Borràs-Comes, Joan, Rafèu Sichel-Bazin, and Pilar Prieto. 2015. "Vocative Intonation Preferences Are Sensitive to Politeness Factors." *Language and Speech* 58 (1): 68–83.

Broe Michael B., and Janet B. Pierrehumbert, eds. 2000. *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. Cambridge: Cambridge University Press.

Browman, Catherine P., and Louis Goldstein. 1992. "Articulatory Phonology: An Overview." *Phonetica* 49:155–180.

Bruce, Gösta. 1977. *Swedish Word Accents in Sentence Perspective*. Lund, Sweden: Gleerup.

Brugos, Alejna, Stefanie Shattuck-Hufnagel, and Nanette Veilleux. 2006. "Transcribing Prosodic Structure of Spoken Utterances with ToBI" (course page). http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/index.htm.

Buring, Daniel. 2013. "Syntax and Prosody, Syntax and Meaning." In *The Cambridge Handbook of Generative Syntax*, edited by Marcel den Dikken, 860–896. Cambridge: Cambridge University Press.

Bryant, Gregory A., and Jean E. Fox Tree. 2005. "Is There an Ironic Tone of Voice?" *Language and Speech* 48:257–277.

Byrd, Dani. 2000. "Articulatory Vowel Lengthening and Coordination at Phrasal Junctures." *Phonetica* 57 (1): 3–16.

Byrd, Dani, Abigail Kaun, Shri Narayanan, and Elliot Saltzman. 2000. "Phrasal Signatures in Articulation." In Broe and Pierrehumbert, 70–87.

Byrd, Dani, and Elliot Saltzman. 1998. "Intragestural Dynamics of Multiple Phrasal Boundaries." *Journal of Phonetics* 26:173–199.

Byrd, Dani, and Elliot Saltzman. 2003. "The Elastic Phrase: Modeling the Dynamics of Boundary-Adjacent Lengthening." *Journal of Phonetics* 31 (2): 149–180.

Calhoun, Sasha. 2010. "The Centrality of Metrical Structure in Signalling Information Structure: A Probabilistic Perspective." *Language* 86 (1): 1–42.

Cambier-Langeveld, Tina, and Alice Turk. 1999. "A Cross-Linguistic Study of Accentual Lengthening: Dutch vs. English." *Journal of Phonetics* 27:171–206.

Campbell, Nick, and Mary E. Beckman. 1997. "Stress, Prominence, and Spectral Tilt." In *Intonation: Theory, Models and Applications: Proceedings of the ESCA Workshop on Intonation*, edited by Antonis Botinis, Georgios Kouroupetroglou, and Georgios Carayannis, 67–70.

Chahal, Dana, and Sam Hellmuth. 2014. "The Intonation of Lebanese and Egyptian Arabic." In Jun 2014b, 365–404.

Chen, Aoju. 2007. "Language-Specificity in the Perception of Continuation Intonation." In Gussenhoven and Riad, 107–141.

Chen, Aoju, Carlos Gussenhoven, and Toni Rietveld. 2004. "Language-Specificity in the Perception of Paralinguistic Intonational Meaning." *Language and Speech* 47:311–349.

Cho, Taehong. 2005. "Prosodic Strengthening and Featural Enhancement: Evidence from Acoustic and Articulatory Realizations of /a, i/ in English." *Journal of the Acoustical Society of America* 117:3867–3878.

Cole, Jennifer. 2000. "Integrating the Phonetics and Phonology of Tone Alignment." In Broe and Pierrehumbert, 168–179.

Cole, Jennifer, and Stefanie Shattuck-Hufnagel. 2016. "New Methods for Prosodic Transcription: Capturing Variability as a Source of Information." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7 (1): 1–29, especially 8.

Connell, Bruce, and Amalia Arvaniti, eds. 1995. *Papers in Laboratory Phonology IV*. Cambridge: Cambridge University Press.

Connell, Bruce, and D. Robert Ladd. 1990. "Aspects of Pitch Realization in Yoruba." *Phonology* 7:1–30.

Dainora, Audra. 2001. "An Empirically Based Probabilistic Model of Intonation in American English." PhD diss., University of Chicago.

Dainora, Audra. 2006. "Modelling Intonation in English: A Probabilistic Approach to Phonological Competence." In Goldstein, Whalen, and Best, 107–132.

Dalton, Martha, and Ailbhe Ní Chasaide. 2007. "Melodic Alignment and Micro-Dialect Variation in Connemara Irish." In Gussenhoven and Riad, 293–315.

Daly, Nicola, and Paul Warren. 2002. "Pitching It Differently in New Zealand English: Speaker Sex and Intonation Patterns." *Journal of Sociolinguistics* 5 (1): 85–96.

Dilley, Laura C., and Meredith Brown. 2007. "Effects of Pitch Range Variation on F0 Extrema in an Imitation Task." *Journal of Phonetics* 35 (4): 523–551.

Dilley, Laura, D. Robert Ladd, and Astrid Schepman. 2005. "Alignment of L and H in Bitonal Accents: Testing Two Hypotheses." *Journal of Phonetics* 33:115–119.

Dilley, Laura, Stefanie Shattuck-Hufnagel, and Mari Ostendorf. 1996. "Glottalization of Word-Initial Vowels as a Function of Prosodic Structure." *Journal of Phonetics* 24:423–444.

D'Imperio, Mariapaola. 2000. "The Role of Perception in Defining Tonal Targets and Their Alignment." PhD diss., Ohio State University.

D'Imperio, Mariapaola. 2001. "Focus and Tonal Structure in Neapolitan Italian." *Speech Communication* 33 (4): 339–356.

D'Imperio, Mariapaola, Gorka Elordieta, Sónia Frota, Pilar Prieto, and Marina Vigário. 2005. "Intonational Phrasing in Romance: The Role of Syntactic and Prosodic Structure." In *Prosodies*, edited by Sónia Frota, Marina Vigário, and M. João Freitas, 59–98. The Hague: Mouton de Gruyter.

D'Imperio, Mariapaola, Jacques Terken, and Michel Piterman. 2000. "Perceived Tone 'Targets' and Pitch Accent Identification in Italian." In *Proceedings of the Australian International Conference on Speech Science and Technology 8*, 201–211.

Elordieta, Gorka, and Nagore Calleja. 2005. "Microvariation in Accentual Alignment in Basque Spanish." *Language and Speech* 48:397–439.

Elordieta, Gorka, and José Ignacio Hualde. 2014. "Intonation in Basque." In Jun 2014b, 405–463.

Féry, Caroline, and Frank Kügler. 2008. "Pitch Accent Scaling on Given, New and Focused Constituents in German." *Journal of Phonetics* 36:680–703.

Fletcher, Janet, Esther Grabe, and P. Warren. 2005. "Intonational Variation in Four Dialects of English: The High Rising Tune." In Jun 2005c, 390–409.

Fougeron, Cécile. 2001. "Articulatory Properties of Initial Segments in Several Prosodic Constituents in French." *Journal of Phonetics* 29:109–135.

Fougeron, Cécile, and Patricia A. Keating. 1997. "Articulatory Strengthening at Edges of Prosodic Domains." *Journal of the Acoustical Society of America* 101:3728–3740.

Fougeron, Cécile, and Sun-Ah Jun. 1998. "Rate Effects on French Intonation: Prosodic Organization." *Journal of Phonetics* 26:45–69.Frota, Sónia. 2002. "Tonal Association and Target Alignment in European Portuguese Nuclear Falls." In *Laboratory Phonology 7*, edited by Carlos Gussenhoven and Natasha Warner, 387–418. Berlin: Mouton de Gruyter.

Frota, Sónia. 2014. "The Intonational Phonology of European Portuguese." In Jun 2014b, 6–424.

Fuchs, Susanne, Caterina Petrone, Amélie Rochet-Capellan, Uwe D. Reichel, and Laura L. Koenig. 2015. "Assessing Respiratory Contributions to F0 Declination in German across Varying Speech Tasks and Respiratory Demands." *Journal of Phonetics* 52:35–45.

Fujisaki, Hiroya. 2004. "Information, Prosody and Modeling with Emphasis on Tonal Features of Speech." In *Proceedings of Speech Prosody,* Nara, Japan, 2004, edited by Bernard Bel and Isabelle Marlien; ISCA Archive. http://www.isca-speech.org/archive/sp2004.

German, James, Janet Pierrehumbert, and Stefan Kaufmann. 2006. "Evidence for Phonological Constraints on Nuclear Accent Placement." *Language* 82 (1): 151–168.

Glasberg, B. R., and Brian Moore 1990. "Derivation of Auditory Filter Shapes from Notched-Noise Data." *Hearing Research* 47:103–138.

Goldsmith, John A. 1976. "Autosegmental Phonology." PhD diss., MIT. Reprinted 1979 as *Outstanding Dissertations in Linguistics*. New York: Garland Publishing.

Goldstein, Louis M., Douglas H. Whalen, and Catherine T. Best, eds. 2006. *Laboratory Phonology 8*. Berlin: Mouton de Gruyter.

Gooden, Shelome. 2014. "Aspects of the Intonational Phonology of Jamaican Creole." In Jun 2014b, 272–301.

Gordon, Matthew. 2005. "Intonational Phonology of Chickasaw." In Jun 2005c, 301–330.

Gordon, Matthew. 2008. "Pitch Accent Timing and Scaling in Chickasaw." *Journal of Phonetics* 36 (3): 521–535.

Grabe, Esther. 1998. “Intonational Phonology: English and German.” PhD diss., Max-Planck-Institute for Psycholinguistics and University of Nijmegen. http://www.phon.ox.ac.uk/files/people/grabe/thesis.html.

Grabe, Esther, Brechtje Post, Francis Nolan, and Kimberley Farrar. 2000. “Pitch Accent Realization in Four Varieties of British English.” *Journal of Phonetics* 28:161–185.

Grabe, Esther, and Paul Warren. 1995. “Stress Shift: Do Speakers Do It or Do Listeners Use It?” In Connell and Arvaniti, 95–110.

Graham, Calbert R. 2014. “Fundamental Frequency Range in Japanese and English: The Case of Simultaneous Bilinguals.” *Phonetica* 71:271–295.

Grice, Martine. 1995. “Leading Tones and Downstep in English.” *Phonology* 12 (2): 183–233.

Grice, Martine, Stefan Baumann, and Ralf Benzmüller. 2005. “German Intonation in Autosegmental-Metrical Phonology.” In Jun 2005c, 55–83.

Grice, Martine, D. Robert Ladd, and Amalia Arvaniti. 2000. “On the Place of ‘Phrase Accents’ in Intonational Phonology.” *Phonology* 17:143–185.

Grice, Martine, Timo Roettger, and Rachid Ridouane. 2015. “Tonal Association in Tashlhiyt Berber: Evidence from Polar Questions and Contrastive Statements.” *Phonology* 32:241–266.

Gussenhoven, Carlos. 1990. “Tonal Association Domains and the Prosodic Hierarchy in English.” In *Studies in the Pronunciation of English: A Commemorative Volume in Honour of A. C. Gimson*, edited by S. M. Ramsaran, 27–37. London: Routledge.

Gussenhoven, Carlos. 1991. “The English Rhythm Rule as an Accent Deletion Rule.” *Phonology* 8:1–35.

Gussenhoven, Carlos. 2000. “The Boundary Tones Are Coming: On the Non-Peripheral Realization of Boundary Tones.” In Broe and Pierrehumbert, 132–151.

Gussenhoven, Carlos. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Gussenhoven, Carlos. 2005. “Transcription of Dutch Intonation.” In Jun 2005c, 118–145.

Gussenhoven, Carlos. 2016. “Analysis of Intonation: The Case of MAE_ToBI.” *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7:1–35.

Gussenhoven, Carlos, and Tomas Riad, eds. 2007. *Tones and Tunes*. Vol. 2, *Experimental Studies in Word and Sentence Prosody*. Berlin: Mouton de Gruyter.

Gussenhoven, Carlos, and A. C. M. Rietveld. 1988. “Fundamental Frequency Declination in Dutch: Testing Three Hypotheses.” *Journal of Phonetics* 16:355–369.

Gussenhoven, Carlos, and Toni Rietveld. 2000. “The Behavior of H* and L* under Variations in Pitch Range in Dutch Rising Contours.” *Language and Speech* 43:183–203.Haspelmath, Martin. 2010. “Comparative Concepts and Descriptive Categories in Crosslinguistic Studies.” *Language* 86 (3): 663–687.

Haspelmath, Martin. 2015. “Defining vs. Diagnosing Linguistic Categories: A Case Study of Clitic Phenomena.” In *How Categorical Are Categories?*, edited by Joanna Błaszczak, Dorota Klimek-Jankowska, and Migdalski Krzysztof, 273–304. Berlin: Mouton de Gruyter.

Herman, Rebecca. 1996. “Final Lowering in Kipare.” *Phonology* 13:171–196.

Herman, Rebecca. 2000. “Phonetic Markers of Global Discourse Structures in English.” *Journal of Phonetics* 28:466–493.

Herman, Rebecca, Mary Beckman, and Kiyoshi Honda. 1996. "Subglottal Pressure and Final Lowering in English." In *Proceedings of the Fourth International Conference on Spoken Language Processing*, edited by H. Timothy Bunnell and William Idsardi, 145–148. IEEE. https://doi.org/DOI: 10.1109/ICSLP.1996.607200.

Hermes, D. J., and J. C. van Gestel. 1991. "The Frequency Scale of Speech Intonation." *Journal of the Acoustical Society of America* 90:97–102.

Hirschberg, Julia, and Gregory Ward. 1992. "The Influence of Pitch Range, Duration, Amplitude and Spectral Features on the Interpretation of the Rise-Fall-Rise Intonation Contour in English." *Journal of Phonetics* 20:241–251.

Hirst, Daniel, Albert Di Cristo, and Robert Espesser. 2000. "Levels of Representation and Levels of Analysis for the Description of Intonation Systems." In *Prosody: Theory and Experiment, Studies Presented to Gösta Bruce*, edited by M. Horne, 51–87. Dordrecht, the Netherlands: Kluwer.

Hyman, Larry M. 2006. "Word-Prosodic Typology." *Phonology* 23:225–257.

Hyman, Larry M. 2009. "How (Not) to Do Phonological Typology: The Case of Pitch Accent." *Language Sciences* 31:213–238.

Jun, Sun-Ah. 2005a. "Korean Intonational Phonology and Prosodic Transcription." In Jun 2005c, 201–229.

Jun, Sun-Ah. 2005b. "Prosodic Typology." In Jun 2005c, 432–458.

Jun, Sun-Ah, ed. 2005c. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, Suh-Ah. 2014a. "Prosodic Typology: By Prominence Type, Word Prosody, and Macro-Rhythm." In Jun 2014b, 520–539.

Jun, Sun-Ah, ed. 2014b. *Prosodic Typology II: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, Sun-Ah, and Janet Fletcher. 2014. "Methodology of Studying Intonation: From Data Collection to Data Analysis." In Jun 2014b, 493–519.

Karlsson, Anastasia M. 2014. "The Intonational Phonology of Mongolian." In Jun 2014b, 187–215.

Katsika, Argyro. 2016. "The Role of Prominence in Determining the Scope of Boundary Lengthening in Greek." *Journal of Phonetics* 55:149–181.

Keane, Elinor. 2014. "The Intonational Phonology of Tamil." In Jun 2014b, 118–153.

Khan, Sameer ud Dowla. 2014. "The Intonational Phonology of Bangladeshi Standard Bengali." In Jun 2014b, 81–117.

Knight, Rachael-Anne. 2008. "The Shape of Nuclear Falls and Their Effect on the Perception of Pitch and Prominence: Peaks vs. Plateaux." *Language and Speech* 51 (3): 223–244.

Knight, Rachael-Anne, and Francis Nolan. 2006. "The Effect of Pitch Span on Intonational Plateaux." *Journal of the International Phonetic Association* 36 (1): 21–38.

Ladd, D. Robert. 1983. "Phonological Features of Intonational Peaks." *Language* 59:721–759.

Ladd, D. Robert 1988. "Declination 'Reset' and the Hierarchical Organization of Utterances." *Journal of the Acoustical Society of America* 84:530–544.

Ladd, D. Robert. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. Robert. 2008. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Ladd, D. Robert. 2011. "Phonetics in Phonology." In *The Handbook of Phonological Theory*, edited by John Goldsmith, Jason Riggle, and Alan C. L. Yu, 348–373. Hoboken, NJ: Blackwell.

Ladd, D. Robert, Ineke Mennen, and Astrid Schepman. 2000. "Phonological Conditioning of Peak Alignment in Rising Pitch Accents in Dutch." *Journal of the Acoustical Society of America* 107:2685–2696.

Ladd, D. Robert, and Astrid Schepman. 2003. "'Sagging Transitions' between High Pitch Accents in English: Experimental Evidence." *Journal of Phonetics* 31:81–112.

Ladd, D. Robert, Astrid Schepman, Laurence White, Louise May Quarmby, and Rebekah Stackhouse. 2009. "Structural and Dialectal Effects on Pitch Peak Alignment in Two Varieties of British English." *Journal of Phonetics* 37 (2): 145–161.

Laniran, Yetunde O., and G. N. Clements. 2003. "Downstep and High Raising: Interacting Factors in Yoruba Tone Production." *Journal of Phonetics* 31:203–250.

Liberman, Mark, and Alan Prince. 1977. "On Stress and Linguistic Rhythm." *Linguistic Inquiry* 8:249–336.

Liberman, Mark Y., and Janet B. Pierrehumbert. 1984. "Intonational Invariance under Changes in Pitch Range and Length." In *Language Sound Structure: Studies in Phonology Presented to Morris Halle*, edited by Mark Aronoff and Richard T. Oehrle, 157–233. Cambridge, MA: MIT Press. Lickley, Robin J., Astrid Schepman, and D. Robert Ladd. 2005. "Alignment of 'Phrase Accent' Lows in Dutch Falling Rising Questions: Theoretical and Methodological Implications." *Language and Speech* 48:157–183.

Maeda, Shinji. 1976. "A Characterization of American English Intonation." PhD diss., MIT.

Menn, Lise, and Suzanne Boyce. 1982. "Fundamental Frequency and Discourse Structure." *Language and Speech* 25:341–383.

Myers, Scott. 2003. "F0 Timing in Kinyarwanda." *Phonetica* 60:71–97.

Nakai, Satsuki, Sari Kunnari, AliceTurk, Kari Suomi, and Riikka Ylitalo. 2009. "Utterance-Final Lengthening and Quantity in Northern Finnish." *Journal of Phonetics* 37:29–45.

Nespor, Marina, and Irene Vogel. 1986. *Prosodic Phonology*. Dordrecht, the Netherlands: Foris.

Niebuhr, Oliver. 2008. "Coding of Intonational Meanings beyond F0: Evidence from Utterance-Final /t/ Aspiration in German." *Journal of the Acoustical Society of America* 124 (2): 1252–1263.

Niebuhr, Oliver. 2012. "At the Edge of Intonation: The Interplay of Utterance-Final F0 Movements and Voiceless Fricative Sounds in German." *Phonetica* 69 (1–2): 7–21.

Nolan, Francis. 2003. "Intonational Equivalence: An Experimental Evaluation of Pitch Scales." In *Proceedings of the 15th International Congress of Phonetic Sciences*, edited by M. J. Solé, D. Recasens, and J. Romero, 771–774.

Nolan, Francis, and Brechtje Post. 2014. "The IViE Corpus." In *The Oxford Handbook of Corpus Phonology*, edited by Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 475–485. Oxford: Oxford University Press.

O'Connor, J. D., and G. F. Arnold. 1973. *Intonation of Colloquial English*. London: Longman.

Ohala, John. 1983. "Cross-Language View of Pitch: An Ethological View." *Phonetica* 40:1–18.

Ostendorf, Mari, Patti Price, and Stefanie Shattuck-Hufnagel. 1996. Boston University Radio Speech Corpus LDC96S36. Philadelphia: Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC96S36.

Peters, Jörg. 2018. "Phonological and Semantic Aspects of German Intonation." *Linguistik Online*. https://doi.org/10.13092/lo.88.4191.

Petrone, Caterina, and Mariapaola D'Imperio. 2011. "From Tones to Tunes: Effects of the F0 Prenuclear Region in the Perception of Neapolitan Statements and Questions." In *Prosodic Categories: Production, Perception and Comprehension*, edited by Sónia Frota, Gorka Elordieta, and Pilar Prieto, 207–230. Berlin: Springer-Verlag.

Petrone, Caterina, and Mariapaola D'Imperio. 2015. "Effects of Syllable Structure on Intonation Identification in Neapolitan Italian." In *Proceedings of the Eighteenth International Congress of Phonetic Sciences*.

Petrone, Caterina, and Oliver Niebuhr. 2014. "On the Intonation in German Intonation Questions: The Role of the Prenuclear Region." *Language and Speech* 57:108–146.

Pierrehumbert, Janet B. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pierrehumbert, Janet B. 1981. "Synthesizing Intonation." *Journal of the Acoustical Society of America* 70:985–995.

Pierrehumbert, Janet B., and Mary Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pierrehumbert, Janet B., and Julia Hirschberg. 1990. "The Meaning of Intonational Contours in the Interpretation of Discourse." In *Intentions in Communication*, edited by Philip R. Cohen, Jerry L. Morgan, and Martha E. Pollack, 271–311. Cambridge MA: MIT Press.

Pike, Kenneth L. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Post, Brechtje. 1999. "Restructured Phonological Phrases in French: Evidence from Clash Resolution." *Linguistics* 37:41–63.

Prieto, Pilar. 2005. "Stability Effects in Tonal Clash Contexts in Catalan." *Journal of Phonetics* 33 (2): 215–242.

Prieto, Pilar. 2007. "Phonological Phrasing in Spanish." In *Optimality-Theoretic Studies in Spanish Phonology*, edited by Fernando Martínez-Gil and Sonia Colina, 39–60. Philadelphia: John Benjamin.

Prieto, Pilar. 2009. "Tonal Alignment Patterns in Catalan Nuclear Falls." *Lingua* 119:865–880.

Prieto, Pilar. 2014. "The Intonational Phonology of Catalan." In Jun 2014b, 43–80.

Prieto, Pilar, Mariapaola D'Imperio, and Barbara Gili Fivela. 2005. "Prenuclear and Nuclear Rising Accents in Romance: Tonal Alignment and Primary and Secondary Associations with Metrical Structure." *Language and Speech* 48 (4): 359–396.

Prieto, Pilar, and José Ignacio Hualde. 2016. "Towards an International Prosodic Alphabet (IPrA)." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7 (1): 5. http://doi.org/10.5334/labphon.11.

Prieto, Pilar, Chilin Shih, and Holy Nibert. 1996. "Pitch Downtrend in Spanish." *Journal of Phonetics* 24:445–473.

Prieto, Pilar, and Francisco Torreira. 2007. "The Segmental Anchoring Hypothesis Revisited: Syllable Structure and Speech Rate Effects on Peak Timing in Spanish." *Journal of Phonetics* 35 (4): 473–500.

Prieto, Pilar, Jan van Santen, and Julia Hirschberg. 1995. "Tonal Alignment Patterns in Spanish." *Journal of Phonetics* 23:429–451.

Prom-On, Santitham, Yi Xu, Wentao Gu, Amalia Arvaniti, Hosung Nam, and Douglas H. Whalen. 2016. "The Common Prosody Platform (CPP): Where theories of prosody can be directly compared." *Proceedings of Speech Prosody 2016*. http://www.isca-speech.org/archive/sp2016/.

Rathcke, Tamara. 2017. "How truncating are 'truncating languages'? Evidence from Russian and German." *Phonetica* 73:194–228.

Redi, Laura. 2003. "Categorical Effects in the Production of Pitch Contours in English." In *Proceedings of the Fifteenth International Congress of Phonetic Sciences,* Barcelona, edited by M. J. Sole, D. Recasens, and J. Romero, 2921–2924; ICPhS Archive. http://www.internationalphoneticassociation.org/icphs/icphs2003.

Remijsen, Bert. 2013. "Tonal Alignment Is Contrastive in Falling Contours in Dinka." *Language* 89 (2): 297–327.

Remijsen, Bert, and Otto Gwado Ayoker. 2014. "Contrastive Tonal Alignment in Falling Contours in Shilluk." *Phonology* 31:435–462.

Rialland, Annie. 2007. "Question Prosody: An African Perspective." In *Tones and Tunes*. Vol. 1, *Typological Studies in Word and Sentence Prosody*, edited by C. Gussenhoven and T. Riad, 35–62. Berlin: Mouton de Gruyter.

Ritchart, Amanda, and Amalia Arvaniti. 2014. "The Form and Use of Uptalk in Southern California English." In *Proceedings of Speech Prosody 7*. https://doi.org/10.21437/SpeechProsody.2016.

Ritter, Simon, and Martine Grice. 2015. "The Role of Tonal Onglides in German Nuclear Pitch Accents." *Language and Speech* 58 (1): 114–128.

Schepman, Astrid, Robin Lickley, and D. Robert Ladd. 2006. "Effects of Vowel Length and 'Right Context' on the Alignment of Dutch Nuclear Accents." *Journal of Phonetics* 34:1–28.

Selkirk, Elizabeth O. 1984. *Phonology and Syntax: The Relationship between Sound and Structure*. Cambridge, MA: MIT Press.

Shattuck-Hufnagel, Stefanie. 1995. "The Importance of Phonological Transcription in Empirical Approaches to 'Stress Shift' versus 'Early Accent': Comments on Grabe and Warren, and Vogel, Bunnell and Hoskins." In Connell and Arvaniti, 128–140.

Shattuck-Hufnagel, Stefanie, Mari Ostendorf, and Ken Ross. 1994. "Stress Shift and Early Pitch Accent Placement in Lexical Items in American English." *Journal of Phonetics* 22:357–388.

Shaw, Jason, A., and Shigeto Kawahara. 2018. "Assessing Surface Phonological Specification through Simulation and Classification of Phonetic Trajectories." *Phonology* 35:481–522.

Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. "ToBI: A Standard for Labelling English Prosody." In *Proceedings of the 1992 International Conference on Spoken Language Processing*; ISCA Archive. http://www.isca-speech.org/archive/icslp_1992.

Silverman, Kim, and Janet Pierrehumbert. 1990. "The Timing of Prenuclear High Accents in English." In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by John Kingston and Mary Beckman, 72–106. Cambridge: Cambridge University Press.

Smiljanić, Rajka. 2006. “Early vs. Late Focus: Pitch-Peak Alignment in Two Dialects of Serbian and Croatian.” In Goldstein, Whalen, and Best, 495–518.

Steedman, Mark. 2014. “The Surface-Compositional Semantics of English Intonation.” *Language* 90 (1): 2–57.

Swerts, Marc, Emiel Krahmer, and Cinzia Avesani. 2002. “Prosodic Marking of Information Status in Dutch and Italian: A Comparative Analysis.” *Journal of Phonetics* 30:629–654.

’t Hart, Johan, René Collier, and Antonie Cohen. 1990. *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Themistocleous, Charalambos. 2014. “Edge-Tone Effects and Prosodic Domain Effects on Final Lengthening.” *Linguistic Variation* 14 (1): 129–160.

Trager, George L., and Henry Lee Smith. 1951. *An Outline of English Structure*. Norman, OK: Battenburg Press.

Truckenbrodt, Hubert. 2002. “Upstep and Embedded Register Levels.” *Phonology* 19:77–120.

Turk, Alice E., and Stefanie Shattuck-Hufnagel. 2000. “Word-Boundary-Related Duration Patterns in English.” *Journal of Phonetics* 28:397–440.

Turk, Alice E., and Stefanie Shattuck-Hufnagel. 2007. “Multiple Targets of Phrase-Final Lengthening in American English Words.” *Journal of Phonetics* 35 (4): 445–472.

Vanrell, Maria del Mar. 2007. “A Tonal Scaling Contrast in Majorcan Catalan Interrogatives.” *Journal of Portuguese Linguistics* 5:147–178.

Venditti, Jennifer. 2005. “The J_ToBI Model of Japanese Intonation.” In Jun 2005c, 172–200.

Vogel, Irene, H. Timothy Bunnell, and Steven Hoskins. 1995. “The Phonology and Phonetics of the Rhythm Rule.” In Connell and Arvaniti, 111–127.

Ward, Gregory, and Julia Hirschberg. 1985. “Implicating Uncertainty: The Pragmatics of Rise-Fall Intonation.” *Language* 61 (4): 747–776.

Warren, Paul. 2015. *Uptalk: The Phenomenon of Rising Intonation*. Cambridge: Cambridge University Press.

Welby, Pauline, and Hélène Loevenbruck. 2006. “Anchored Down in Anchorage: Syllable structure and Segmental Anchoring in French.” *Italian Journal of Linguistics* 18:74–124.

Wong, Wai Yi P., Marjorie K. M. Chan, and Mary E. Beckman. 2005. “An Autosegmental-Metrical Analysis and Prosodic Annotation Conventions for Cantonese.” In Jun 2005c, 271–300.

Xu, Yi. 2005. “Speech Melody as Articulatorily Implemented Communicative Functions.” *Speech Communication* 46 (3–4): 220–251.

Xu, Yi, Albert Lee, Santitham Prom-on, and Fang Liu. 2015. “Explaining the PENTA Model: A Reply to Arvaniti and Ladd (2009).” *Phonology* 32 (3): 505–535.

Yuen, Ivan. 2007. “Declination and Tone Perception in Cantonese.” In Gussenhoven and Riad, 63–77.

Commentary on Chapter 1: Introducing Flexibility into Autosegmental-Metrical Phonology

Martine Grice

Introduction

In this chapter, Amalia Arvaniti provides a comprehensive overview of the autosegmental-metrical (AM) approach to intonational phonology, tracing its development over the last three decades. Although there is some diversity among models that come under the AM umbrella, not least for the analysis of the intonation of English, this is currently the most widespread approach to the study of intonation, and it continues to be employed for analyzing a broad range of well-researched and lesser-known languages (Jun 2005, 2014).

Main Advantage of AM: Separation of Tune and Text

The main advantage of the AM approach over previous and competing ones is that the *tune*, in the form of H and L tones, is represented on a tier that is autonomous with respect to the *text*. This autonomy allows for generalizations to be made about tunes that occur across different words and phrases. Because the H and L tones need only to be specified for certain privileged syllables in an utterance, the generalization of patterns to new examples of different lengths is relatively straightforward. Thus the texts in (1) and (2) can be produced with what is essentially the same intonation pattern, represented on the tone tier by an identical sequence of tones, a sequence of L, H, and L, representing the rise to a pitch peak followed by a fall to low pitch.

(1) LHL
Right
(2) L H L
I **agree** with you

Both the tune and the text tiers have some form of structure, which plays a role in determining how they are synchronized. We turn to the structuring of the two tiers.

Structure of Tiers and Complex Intonational Primitives

Text

On the text tier, words are analyzed into smaller constituents, such as syllables and feet, and at the same time, they are grouped together into larger constituents, reflecting phonological phrasing. These constituents are organized hierarchically, with designated heads. The general mechanism for synchronizing the two tiers is clear: tones are associated with tone-bearing units (TBUs) via association lines, and because tones are relatively sparsely distributed, reference has to be made to the structure of the text tier to determine the TBUs, in particular stressed syllables (Goldsmith 1979; Liberman 1979; Leben 1976). Taking the structure of the text into account, association to a stressed syllable implies association to a foot or prosodic word of which this syllable is the head (Beckman and Pierrehumbert 1986; Pierrehumbert and Beckman 1988). There can also be association to the edges of constituents, and in some languages (Korean, Turkish and Hindi, for example), this is the only association possible (Jun 2005, 2014; Féry 2017). Association to edges often means that the tone is placed at or near a particular edge, but it can also have a secondary association with a TBU close to the edge (Pierrehumbert and Beckman 1988; Grice, Ladd, and Arvaniti 2000; inter alia). This means that the tone is anchored to a designated syllable, for example, the final or penultimate syllable in the phrase, or in some cases even a lexically stressed syllable. Not all languages appear to have such secondary associations, and not all models in the AM framework allow for this possibility.

Tune

The structure of the tone tier, representing the tune, consists of H and L tones, which are generally agreed to be intonational primitives. These tones can be combined in different ways to form more complex structures: either they are grouped together within the tone tier, with a structure permitting one tone to be strong (represented using the star notation), or they are multiply associated with a TBU without a strength relation. They form two types of structures, which are generally taken to be building blocks of intonational tunes, which Arvaniti refers to as complex primitives: pitch accents and boundary tones. These are usually defined according to not only (i) their association properties, that is, whether they associate with heads or edges of constituents, but also (ii) their function within the system, that is, whether they are prominence-cueing or demarcative.

Cross-linguistic preferences for relating association properties of tones to their function have led to overoptimization in AM models toward the typical case. This can be a straitjacket for the analysis of new languages and even for the analysis of new phenomena within languages already under investigation. Allowing for tones to have associations to a constituent in the prosodic structure, without necessarily further specifying head or edge association, in addition to separating the association properties from the function, provides a more powerful and flexible tool for the analysis of a wider set of phenomena. This is especially important, given that we still know little about the prosodic systems of the majority of the world's languages.

Two examples illustrate the need for flexibility in AM models of intonation: Tashlhiyt Berber has prominence-cueing tones in the absence of a stable metrical structure and hence the absence of a clearly defined head, and Maltese has tones that appear to function similarly to the regular pitch accents in the language, but these tones are at a word edge rather than its head.

In Tashlhiyt Berber, prominence-cueing tones are highly variable in their position, reflected in F0 peak alignment (Grice, Ridouane, and Roettger 2015; Roettger and Grice 2015; Roettger 2017). In addition to variability in identical pragmatic contexts (see figure 1c.1), some of the variability has been shown to be affected by factors such as sonority, syllable weight, and sentence modality, as well as speaker-specific tendencies. Such variability precludes an analysis whereby tones are associated to specific syllables. Rather, the general picture indicates the possibility of an association to a larger constituent, with probabilistic distributions reflecting their exact placement. Sentence modality plays a role insofar as a greater number of peaks are found on the final syllable in polar questions than in contrastive statements.

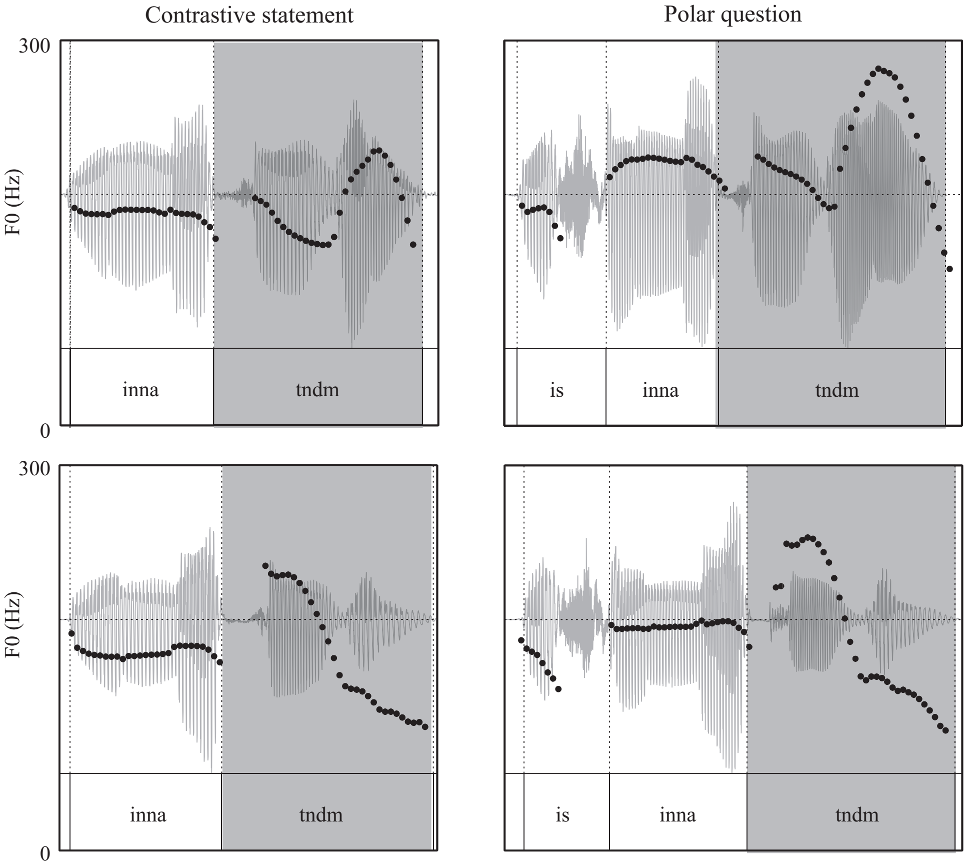


Figure 1c.1
Tashlhiyt Berber. Examples of final peaks (top) and penultimate peaks (bottom) in Tashlhiyt in statements with contrastive focus (left)—*inna tndm* (He said: "She regretted")—and questions (right)—*is inna tndm* (Did he say "she regretted"?). Shaded area indicates the word on which the peak occurs. *Source*: Adapted from Grice, Ridouane, and Roettger (2015).

In phrase-initial position, an H tone can be realized as a peak with variable alignment within a constituent or optionally as a plateau extending over the whole constituent (Bruggeman, Roettger, and Grice 2017; Bruggeman 2018; see figure 1c.2). This further supports an analysis of tones having an association to a constituent without this being reflected in alignment with either a head or an edge. Thus, although the tone serves the function of cueing prominence (e.g., by virtue of occurring on the focused constituent, or being used in explicit contrast), it does not have the association properties of a regular pitch accent, that is, association to a head. Allowing for association of tones to a constituent without them percolating to the head or edge of this constituent may be a way forward for analyzing further languages that—like Tashlhiyt—lack lexical stress.

Maltese poses a further challenge to the current understanding of tune-text association in the form of an H tone that occurs at the beginning of words in certain pragmatic contexts, such as in questions, imperatives, and vocatives (Vella 2007). The complex question word in *ma' min minnhom* (With which one of them?) in figure 1c.3 has lexical stress on the penultimate syllable, following the stress rules of Maltese. In direct

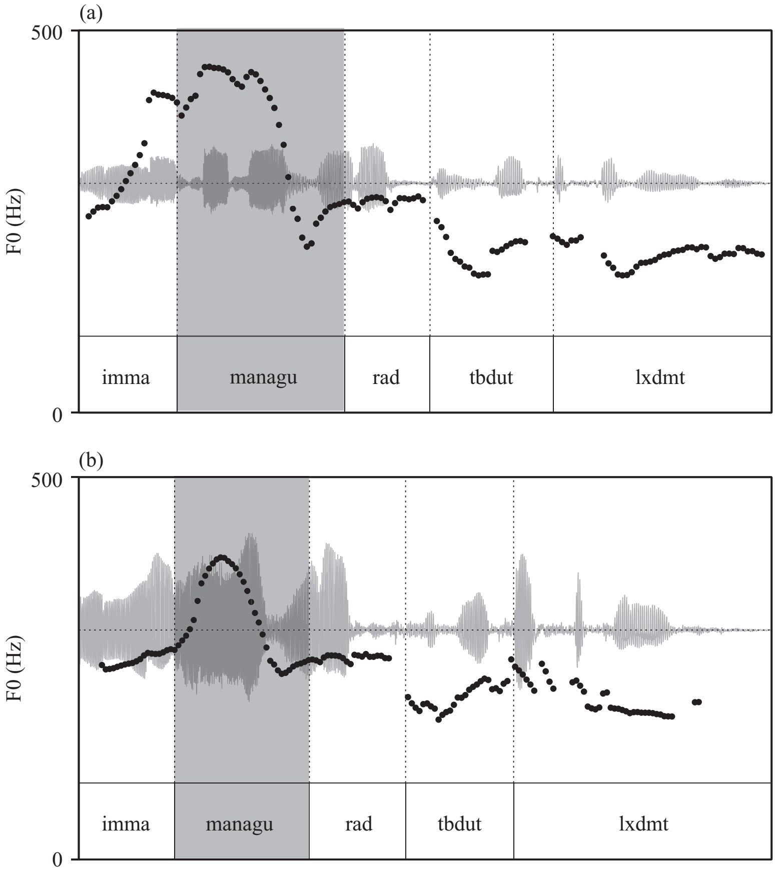


Figure 1c.2
Tashlhiyt Berber. Example of (a) plateau and (b) peak on question word constituent *managu* (when) in *Imma managu rad tbdut lxdmt?* (So when will you start work?). *Source*: Adapted from Bruggeman, Roettger, and Grice (2017).

questions (the top panel of figure 1c.3), there is a fall in pitch on the first syllable, *ma'* (note that the apostrophe is part of the orthography and does not reflect stress). In indirect questions (the bottom panel of figure 1c.3), by contrast, the fall occurs on the stressed syllable. This pattern, in which the fall in pitch is on the first syllable in direct questions and on the stressed syllable in indirect questions, was found to be stable across speakers in a production study involving scripted dialogues (Grice, Vella, and Bruggeman 2019).

Thus Maltese employs a word-initial tone in direct questions and a tone associating with the lexical stress—in line with other pitch accents in the language—in indirect questions. The word-initial tone cannot be analyzed as marking the edge of a larger prosodic constituent (intermediate phrase/intonational phrase), because question words can occur in medial and final positions in the phrase and have word-initial tones in these cases too. Neither is there evidence of a smaller, phrase-internal boundary before the question word. The language thus appears to make use of both edges and heads of words for the association of tones, although there is no evidence for one association leading to a demarcative function and the other to a prominence-cueing one.

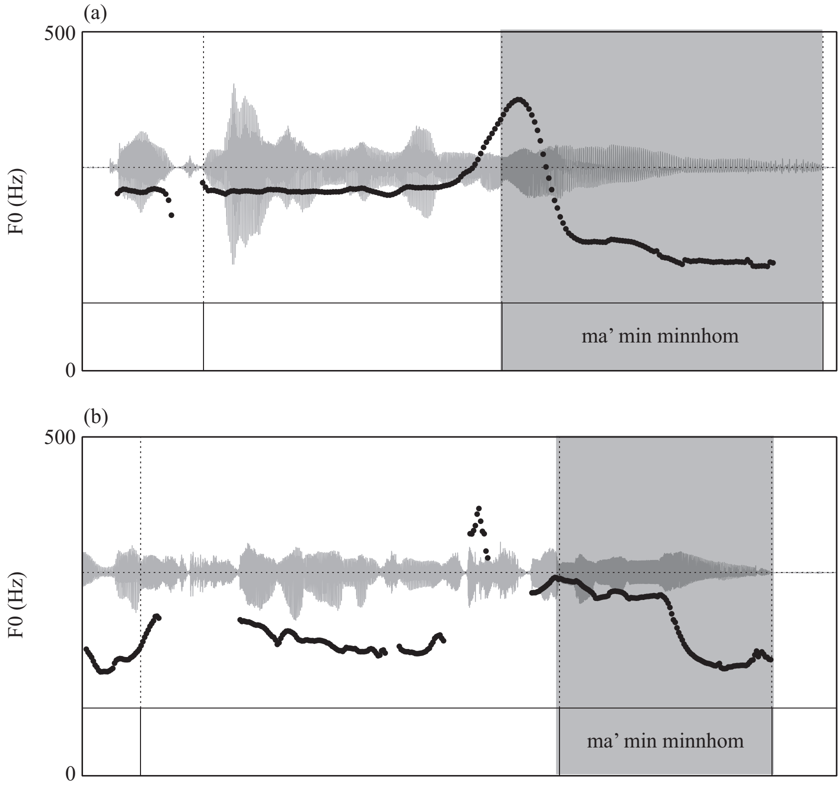


Figure 1c.3
Maltese. F0 contours for a direct question (top) *Mar jgħum ir-Ramla ma’ min minnhom?* (With which one of them did he go swimming at Ramla?) and for an indirect question (bottom) *Il-mistoqsija li għandna bżonn insaqsu hija: ma’ min minnhom* (The question we need to ask is: With which one of them?). Shaded area indicates the complex question word, *ma’min minnhom* (With which one of them?). *Source*: Adapted from Grice, Vella, and Bruggeman (2019).

The reported research on both Tashlhiyt Berber and Maltese clearly calls for a separation of association properties (head, edge, higher-level constituent without percolation) and their function (prominence-cueing, demarcative) when determining the intonational inventory of a language.

Making Generalizations in AM: Beyond Pitch Accents and Edge Tones

As pointed out by Arvaniti, a successful model of intonation should provide a principled way of accounting for differences among the contours of a language and be able to make generalizations, such that it can capture instantiations of the “same” melody across different segmental strings, that is, the same tune on different texts. Although there is evidence for the compositionality of *tunes* (e.g., Pierrehumbert and Hirschberg 1990), meanings being assigned to particular pitch accents or edge tones separately, the general practice is to take into account particular combinations of pitch accents and edge tones. For English and other West Germanic languages, this is often the nuclear tune—the nuclear pitch accent (often equated with the final pitch accent in the phrase) and the following edge tones. For instance, the intonation of examples (1) and (2) are both analyzable as L + H* L-L%. A longer phrase such as in example (3) can be considered to be "the same," even if there are two accents, because it is the final accent that is generally considered to bear the main pragmatic meaning of a phrase-length tune, although there is a growing interest within the AM framework in the meaning of initial (secondary) accents (Braun 2006; Féry and Kügler 2008; Petrone and Niebuhr 2014; Bishop 2013, 2017) and phrase-length patterns (Welby 2003; Baumann and Rathcke 2013; Baumann, Mertens, and Kalbertodt 2017).

```
           H*          L + H*   L-L%
(3)  I think you're right.
```

The concept of nuclear accent was developed with West Germanic in mind—languages that are flexible in their placement of accents. In these languages, the nuclear accent on a focused constituent can be followed by a number of words that are deaccented (they are not accented even if they are in principle accentable, e.g., nouns). In languages that do not make such widespread use of deaccenting after a focal accent, the primacy of the final pitch accent has long been called into question. Languages that tend to accent more densely may not assign any special status to the final accent in a phrase. This is the case for varieties of Arabic (Hellmuth 2014, forthcoming) and for many Romance languages (Swerts, Krahmer, and Avesani 2002; Frota and Prieto 2015), as well as for Greek (as discussed by Arvaniti).

Torreira and Grice (2018) discuss a number of tunes in Spanish that not only call into question the importance of final pitch accents but also call for a different way to describe the similarity of tunes across utterances of different lengths. They show that the tunes they investigate are better conceived of as tones that are associated differently depending on the metrical structure of the phrase. They argue that the association properties of these tones depend on the availability of suitable heads in the metrical structure. The evidence put forward stems from two sources: conversational data in which the focus structure and pragmatic meaning can be inferred from the context and experimental data in an imitation and completion task. In the Spanish low-rise-fall contour, as shown in figure 1c.4, they provide evidence for differential association of the H tone, depending on whether the utterance consists of two or more prosodic words or only one. In the former case, H has a head association, in the latter case it has an edge association. By contrast, the first L always has a head association, as shown in (4), where it is associated with the stressed syllable of *amigo* in (4a) and with the stressed syllable of *Manolo* in (4b).

```
               L              H  L               L   H (L)
(4)  a.  El amigo de Manolo        b.  Manolo
```

The common practice of describing tunes as consisting of nuclear pitch accents and edge tones misses this generalization, as does an analysis of tones with a fixed association type (e.g., a tune containing H*). Torreira and Grice propose a clearer separation between tones and metrical structure, in which association can result from melody-specific principles of tune-text association. They thus propose a more abstract notion of tune, for example, representing both (4a) and (4b) as LHL, with a priority association of the first L to the first stressed syllable in the phrase, ensuring the presence of L* and variable association of the following H and L tones. Other tunes in other languages may have different priority associations. Again, allowing for this flexibility would make it easier to capture similarities in other languages that may be lost if considering tunes as strictly composed of (nuclear) pitch accents and edge tones.

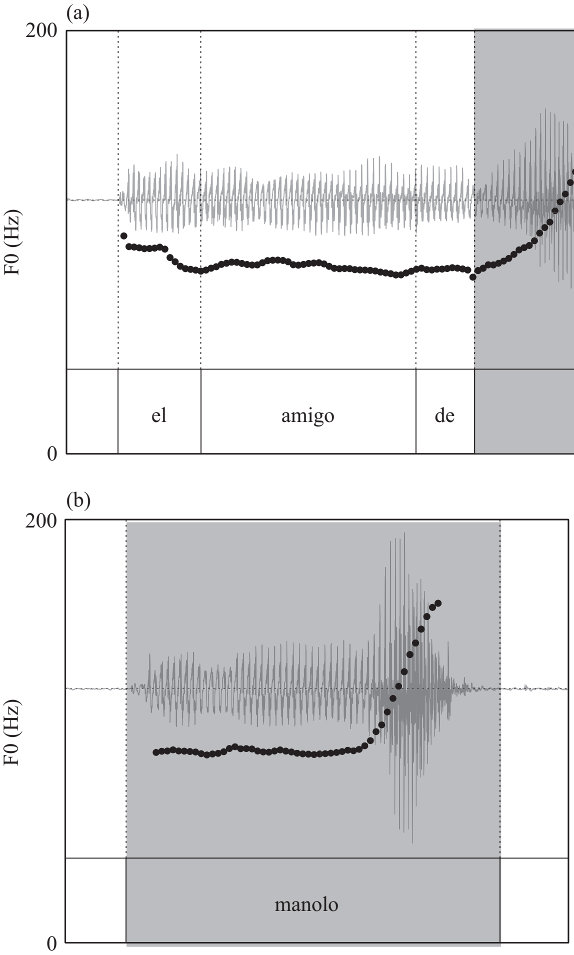


Figure 1c.4
Spanish. F0 contours and speech waveforms with intonation contours that are "the same" on (a) two prosodic words, *el amigo de Manolo* (the friend of Manolo), and (b) one prosodic word, *Manolo* (Manolo). *Source*: Adapted from Torreira and Grice (2018).

One aspect of this Spanish contour needs clarification: the final pitch does not fall when there is only one prosodic word, indicated informally in (4b) by parentheses around L, but it does fall when there are two or more prosodic words in the phrase. The former case is interpreted as truncation, as discussed at length for other languages by Arvaniti, and put into a larger context of tune-text adjustments.

Tune-Text Adjustments

Truncation of pitch contours involves adjustments to the tune when more than one tone is crowded onto a TBU. Another possible adjustment in the case of tonal crowding is to make room for the realization of a particular tune by extending the text. This can entail lengthening of vowels (Grice, Savino, and Refice 1997), the addition of vowels

(Grice, Savino, and Roettger 2018), or blocking of vowel devoicing (Kilbourn-Ceron and Sonderegger 2017), all serving to facilitate the full realization of a complex contour.

Both types of adjustment—to the *tune*, involving tonal identity, and to the *text*, involving segmental properties of the TBUs—may be characterized by continuous parameters, lending themselves to a phonetic implementation account, as discussed by Arvaniti. However, not all adjustments to the tune can be accounted for as phonetic. The case of *Manolo* in Spanish in figure 1.c4a is a case in point. There is in fact ample time for a rise-fall in pitch to occur on the final open syllable, there being, for example, no final voiceless consonant to shorten the voiced portion in the syllable. Typical cases of truncation attributed to phonetic implementation have such a voiceless coda (see, e.g., Grabe 1998 for German and Grabe, Post, Nolan, and Farrar 2000 for varieties of English). However, Rathcke (2016) has shown that patterns of truncation are language-specific, as well as dependent on the association properties of a particular tone. Furthermore, she discusses the truncation of the fall in Russian rise-falls as "a categorical abandonment of an underlying boundary tone" (223). This is clearly a phonological analysis, similar to Grice's (1995) analysis of a similar fall in Palermo Italian in which truncation is treated in terms of the lack of association to a TBU.

Adjustments to the text are not always attributable to phonetic implementation either. Vowels that are inserted are sometimes as long as or even longer and more sonorous than lexical vowels (for Tashlhiyt, see Roettger 2017; for Italian, see Grice, Savino, and Roettger 2018). Moreover, durational adjustments are sometimes attributed to phonological lengthening, represented as an extra metrical beat (Selkirk 1984) or an extra mora (see for instance Strunk 1983 on vowels with an added mora in Sanskrit).

Any phonological account of these adjustments compromises the independence of tune and text, one of the basic tenets of AM theory (see Roettger and Grice 2019 for further discussion).

Problems with Categorization in AM

Although AM models are widely used as a standard for the transcription of intonation (Beckmann, Hirschberg, and Shattuck-Hufnagel 2005), there are still considerable problems of intertranscriber consistency (Pitrelli, Beckman, and Hirschberg 1994; Grice et al. 1996; Mayo, Aylett, and Ladd 1997; Jun et al. 2000; and Escudero et al. 2012, inter alia). An AM analysis of an utterance not only determines the association properties and functions of tones; it also determines the tonal makeup of the tune, that is, which tones and sequences of tones are present. For example, in (3) the prenuclear pitch accent is H* rather than L+H*, and the nuclear pitch accent is L+H* rather than H* or L*+H. Such paradigmatic aspects of the analysis are particularly prone to intertranscriber inconsistencies.

This is in part because of the complex relationship between phonetic parameters and abstract AM categories, involving multiple dimensions. Thus, as Arvaniti points out, peak alignment and scaling are not sufficient for capturing distinctions between these categories. In fact, even within one language community, individual speakers and listeners use these dimensions in different ways and to varying degrees to produce and perceive one and the same category (Grice et al. 2017; Cangemi, Krüger, and Grice 2015 for German; see similar results for English in Burdin, Holliday, and Reed 2018). Thus, where one speaker or listener will pay more attention to one dimension, for example, the alignment of F0 peaks, another will focus on a different one, for example, F0 scaling. These differences might be captured by a different weighting of cues or, at least partially, by measures that collapse various dimensions over a longer stretch than a simple peak position, such as the tonal center of gravity (Barnes et al. 2012).

Concentrating on one parameter only, for example, alignment of peaks, can lead to the conclusion that a language has no pitch-accent distinction, such as in the case of Egyptian Arabic, because it has been shown that different speakers make use of different phonetic parameters to express the same contrast (Cangemi et al. 2016). Bruggeman et al. (2018) showed that a measure of the tonal center of gravity was able to capture important distinctions across different speakers that would otherwise be lost if relying on individual F0 peak properties as a single cue.

The nature of intonational meaning—with functions ranging from marking information structure, managing discourse, and encoding attitudinal and indexical aspects of communication—makes it difficult to establish how far the parameter manipulations discussed here are equivalent in their function. Despite problems in pinning down the relative importance of different parameters, the necessity to have labels for abstract entities has meant that the AM models are still prevalent and are still the models of choice for embarking on an analysis of a new language. Allowing for flexibility in these models will ensure that the model will be able to accommodate and account for the considerable range of variation within and across languages that we are beginning to discover.

References

Barnes, Jonathan A., Nanette Veilleux, Alejna Brugos, and Stefanie Shattuck-Hufnagel. 2012. "Tonal Center of Gravity: A Global Approach to Tonal Implementation in a Level-Based Intonational Phonology." *Laboratory Phonology* 3 (2): 337–383.

Baumann, Stefan, Jane Mertens, and Janina Kalbertodt. 2017. "The Influence of Information Status on Prenuclear Accents." Poster presentation at Phonetics and Phonology in Europe (PaPE), Cologne, June.

Baumann, Stefan, and Tamara Rathcke. 2013. "Disambiguating the Scope of Negation by Prosodic Cues in Three Varieties of German." *Lingua* 131:29–48.

Beckman, Mary E., Julia Hirschberg, and Stefanie Shattuck-Hufnagel. 2005. "The Original ToBI System and the Evolution of the ToBI Framework in Jun (2005)." In *Prosodic Models and Transcription: Towards Prosodic Typology*, edited by S-A. Jun, 9–54. Oxford: Oxford University Press.

Beckman, Mary E., and Janet B. Pierrehumbert. 1986. "Intonational Structure in Japanese and English." *Phonology Yearbook* 3:255–309.

Bishop, Jason. 2013. "Prenuclear Prominence: Phonetics, Phonology, and Information Structure." PhD diss., UCLA.

Bishop, Jason. 2017. "Focus Projection and Prenuclear Accents: Evidence from On-Line Lexical Processing." *Language, Cognition and Neuroscience* 32 (2): 236–253.

Braun, Bettina. 2006. "Phonetics and Phonology of Thematic Contrast in German." *Language and Speech* 49 (4): 451–493.

Bruggeman, Anna. 2018. "Lexical and Postlexical Prominence in Tashlhiyt Berber and Moroccan Arabic." PhD diss., University of Cologne. https://kups.ub.uni-koeln.de/11189/.

Bruggeman, Anna, Francesco Cangemi, Simon Wehrle, Dina El Zarka, and Martine Grice. 2018. "Unifying Speaker Variability with the Tonal Centre of Gravity." In *Proceedings of the Conference on Phonetics and Phonology in German-Speaking Countries*, edited by Belz, Malte, Christine Mooshammer, Susanne Fuchs, Oksana Rasskazova, and Marzena Żygis. https://edoc.hu-berlin.de/handle/18452/19531.

Bruggeman, Anna, Timo B. Roettger, and Martine Grice. 2017. "Question Word Intonation in Tashlhiyt Berber: Is 'High' Good Enough?" *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8 (1): 1–17.

Burdin, Rachel, Nicole Holliday, and Paul Reed. 2018. "Rising above the Standard: Variation in L+H* Contour Use across Five Varieties of American English." In *Proceedings of Speech Prosody 9*, edited by Katarzyna Klessa, Jolanta Bachan, Agnieszka Wagner, Maciej Karpiński, and Daniel Śledziński, 354–358; ISCA Archive. https://www.isca-speech.org/archive/SpeechProsody.

Cangemi Francesco, Martina Krüger, and Martine Grice. 2015. "Listener-Specific Perception of Speaker-Specific Productions in Intonation." In *Individual Differences in Speech Production and Perception*, edited by S. Fuchs, D. Pape, C. Petrone, and P. Perrier, 123–145. Frankfurt: Peter Lang.

Cangemi, Francesco, Dina El Zarka, Simon Wehrle, Stefan Baumann, and Martine Grice. 2016. "Speaker-Specific Intonational Marking of Narrow Focus in Egyptian Arabic." In *Proceedings of Speech Prosody 8*, edited by Jonathan Barnes, Alejna Brugos, Stefanie Shattuck-Hufnagel, and Nanette Veilleux, 335–339; ISCA Archive. https://www.isca-speech.org/archive/SpeechProsody_2016/.

Escudero, David, Lourdes Aguilar, Maria del Mar Vanrell, and Pilar Prieto. 2012. "Analysis of Inter-Transcriber Consistency in the Cat ToBI Prosodic Labelling System." *Speech Communication* 54 (4): 566–582.

Féry, Caroline. 2017. *Intonation and Prosodic Structure*. Key Topics in Phonology. Cambridge: Cambridge University Press.

Féry, Caroline, and Frank Kügler. 2008. "Pitch Accent Scaling on Given, New and Focused Constituents in German. *Journal of Phonetics* 36 (4): 680–703.

Frota, Sónia, and Pilar Prieto. 2015. "Intonation in Romance: Systemic Similarities and Differences." In *Intonation in Romance*, edited by S. Frota and P. Prieto, 392–418. Oxford: Oxford University Press.

Goldsmith, John A. 1979. *Autosegmental Phonology*. New York: Garland Press.

Grabe, Esther. 1998. "Pitch Accent Realization in English and German." *Journal of Phonetics* 26(2): 129–143.

Grabe, Esther, Brechtje Post, Francis Nolan, and Kimberley Farrar. 2000. "Pitch Accent Realization in Four Varieties of British English." *Journal of Phonetics* 28(2): 161–185.

Grice, Martine. 1995. *The Intonation of Interrogation in Palermo Italian: Implications for Intonation Theory*. Tübingen: Niemeyer.

Grice, Martine, D. Robert Ladd, and Amalia Arvaniti. 2000. "On the Place of Phrase Accents in Intonational Phonology." *Phonology* 17 (2): 143–185.

Grice, Martine, Matthias Reyelt, Ralf Benzmuller, Jörg Mayer, and Anton Batliner. 1996. "Consistency in Transcription and Labeling of German Intonation with GToBI." In *Proceedings of the International Conference on Spoken Language Processing,* 1716–1719; ISCA Archive. https://www.isca-speech.org/archive/icslp_1996/i96.

Grice, Martine, Rachid Ridouane, and Timo B. Roettger. 2015. "Tonal Association in Tashlhiyt Berber: Evidence from Polar Questions and Contrastive Statements." *Phonology* 32:241–266.

Grice, Martine, Simon Ritter, Henrik Niemann, and Timo B. Roettger. 2017. “Integrating the Discreteness and Continuity of Intonational Categories.” *Journal of Phonetics* 64 (1): 90–107.

Grice, Martine, Michelina Savino, and Mario Refice. 1997. “The Intonation of Questions in Bari Italian: Do Speakers Replicate Their Spontaneous Speech When Reading?” *Phonus* 3:1–7.

Grice, Martine, Michelina Savino, and Timo B. Roettger. 2018. “Word Final Schwa Is Driven by Intonation: The Case of Bari Italian.” *Journal of the Acoustical Society of America* 143 (4): 2474–2486.

Grice, Martine, Alexandra Vella, and Anna Bruggeman. 2019. “Stress, Pitch Accent, and Beyond: Intonation in Maltese Questions.” *Journal of Phonetics* 76:1–17.

Hellmuth, Sam. 2014. “Investigating Variation in Arabic Intonation: The Case for a Multi-Level Corpus Approach. In *Perspectives on Arabic Linguistics XXIV–XXV*, edited by Samira Farwaneh and Hamid Ouali, 63–89. Philadelphia: John Benjamin.

Hellmuth, Sam. Forthcoming. “Text-Tune Alignment in Tunisian Arabic Yes-No Questions.” In *Prosodic Variation (with)in Languages: Intonation, Phrasing and Segments*,” edited by M. Cruz, S. Frota, and P. Oliveira. Sheffield, UK: Equinox.

Jun, Sun-Ah, ed. 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, Sun-Ah, ed. 2014. *Prosodic Typology II: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, Sun-Ah, Sook-hyang Lee, Keeho Kim, and Yong-Ju Lee. 2000. “Labeler Agreement in Transcribing Korean Intonation with K-ToBI.” In *Proceedings of the Sixth International Conference on Spoken Language Processing*, 211–214; ISCA Archive. http://www.isca-speech.org/archive/icslp_2000.

Kilbourn-Ceron, Oriana, and Morgan Sonderegger. 2017. “Boundary Phenomena and Variability in Japanese High Vowel Devoicing.” *Natural Language and Linguistic Theory* 36 (1): 1–43.

Leben, William R. 1976. “The Tones in English Intonation.” *Linguistic Analysis* 2:49–107.

Liberman, Mark. 1979. *The Intonational System of English*. New York: Garland Press.

Mayo, Catherine, Matthew Aylett, and D. Robert Ladd. 1997. “Prosodic Transcription of Glasgow English: An Evaluation Study of GlaToBI.” In *Intonation: Theory, Models and Applications* (*Proceedings of an ESCA Workshop*, September 18–20, 1997), edited by Antonis Botinis, Georgios Kouroupetroglou, and George Carayiannis, 231–234. Athens, Greecee: ESCA and University of Athens Department of Informatics.

Petrone Caterina, and Oliver Niebuhr. 2014. “On the Intonation of German Intonation Questions: The Role of the Prenuclear Region.” *Language and Speech* 57 (1): 108–146.

Pierrehumbert, Janet B., and Mary E. Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pierrehumbert, Janet, and Julia Hirschberg. 1990. “The Meaning of Intonation Contours in the Interpretation of Discourse.” In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–310. Cambridge, MA: MIT Press.

Pitrelli, John, Mary E. Beckman, and Julia Hirschberg. 1994. “Evaluation of Prosodic Transcription Labelling Reliability in the ToBI Framework.” In *Proceedings of the Third International Conference on Spoken Language Processing*, 123–126; ISCA Archive. http://www.isca-speech.org/archive/icslp_1994.

Rathcke, Tamara. 2016. “How Truncating Are ‘Truncating Languages’? Evidence from Russian and German.” *Phonetica* 73:194–228.

Roettger, Timo B. 2017. *Tonal Placement in Tashlhiyt: How an Intonation System Accommodates to Adverse Phonological Environments*. Berlin: Language Science Press.

Roettger, Timo B., and Martine Grice. 2015. “The Role of High Pitch in Tashlhiyt Tamazight (Berber): Evidence from Production and Perception.” *Journal of Phonetics* 51:36–49.

Roettger, Timo B., and Martine Grice. 2019. “The Tune Drives the Text: Competing Information Channels of Speech Shape Phonological Systems.” *Language Dynamics and Change* 9:265–298.

Selkirk, Elisabeth. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Strunk, Klaus. 1983. *Typische Merkmale von Fragesätzen und die altindische “Pluti.”* Munich: Verlag der Bayerischen Akademie der Wissenschaften.

Swerts, Marc, Emiel Krahmer, and Cinzia Avesani. 2002. “Prosodic Marking of Information Status in Dutch and Italian: A Comparative Analysis.” *Journal of Phonetics* 30 (4): 629–654.

Torreira, Francisco, and Martine Grice. 2018. “Melodic Constructions in Spanish: Metrical Structure Determines the Association Properties of Intonational Tones.” *Journal of the International Phonetic Association* 48 (1): 9–32.

Vella, Sandra. 2007. “The Phonetics and Phonology of wh-Question Intonation in Maltese.” In *Proceedings of the International Congress of Phonetic Sciences XVI*, edited by William J. Barry and Jürgen Trouvain, 1285–1289. Saarbrücken: University of the Saarland.

Welby, Pauline. 2003. “Effects of Pitch Accent Position, Type, and Status on Focus Projection.” *Language and Speech* 46 (1): 53–81.

Author Response to the Commentary: Prosodic Typology and the Handling of Variability in Intonation

Amalia Arvaniti

Grice raises several interesting points in her commentary. Many of them are the result of newer research that breaks with some traditional approaches and practices within the autosegmental-metrical (AM) approach to intonational phonology. This break from tradition is clearly beneficial not only for AM but for understanding intonation and its functions independently of theoretical framework adopted.

In this response, I focus on two points that emerge from Grice's commentary: prosodic typology and the treatment of variability in the study of intonation. I address these two points in turn and provide some suggestions for a way forward, including research that I think is likely to lead to new insights and advancements in the study of intonation.

Prosodic Typology

Grice uses data from Berber and Maltese to question the distinction between tones with a demarcative function (*edge tones* in AM)—tones that associate with prosodic constituent edges—and prominence lending tones (*pitch accents* in AM)—tones that associate with prosodic heads. Grice questions this binary distinction that conflates association with function and provides evidence that tones can show substantial variability in their synchronization with the segmental string. She thus proposes "a separation of association properties (head, edge, higher-level constituent) and their function (prominence-cueing, demarcative)" (see "Structure of Tiers and Complex Intonational Primitives").

In my view, Grice's suggestion opens up new possibilities for refining prosodic typology. First, I concur with her use of the term *prominence-cueing*, following the suggestion made in Ladd (2008, 2): pitch accents are not prominence-lending but prominence-cueing (a term coined by Francis Nolan). In other words, the function of some tones is to alert the listener to the fact that a given constituent is already prominent (due to its strength in the prosodic hierarchy). Second, I follow the proposal of Pierrehumbert and Beckman (1988), according to which tones associate to prosodic constituents rather than directly to stressed syllables. For example, a tone may associate to an intermediate phrase and percolate down to the phrase's head, which will be the stressed syllable of the metrically strongest constituent in the phrase (or designated terminal element, DTE for short). Thus, what Pierrehumbert and Beckman envision is not a direct relation between tones and tone-bearing units (TBUs) but an indirect one, while Grice seems to suggest a direct relationship between tones and TBUs and indirect connection to higher levels of the prosodic hierarchy.

If we follow these two proposals, it is possible to reframe Grice's suggestion as a consistent and testable typological distinction between languages in which tones percolate all the way down to heads of feet (i.e., lexically stressed syllables or DTEs) versus languages in which tones associate to specific prosodic constituents, such as the prosodic word or the intermediate phrase, but without this association percolating down to the constituent's head, either because there is not one, as in Berber, or because cueing head prominence by localizing it to a DTE is not critical, as in Maltese. Another way to understand this last point is to say that the Maltese tonal events discussed by Grice cue word prominence but not the prominence of a specific syllable.

This reanalysis allows us to capture attested differences between languages with and without lexical stress but also differences between the latter. For example, Korean is a language without lexical stress in which tones associate to phrases, the accentual and intonational phrase. This association percolates to constituent edges, and thus tones are realized in the periphery of their constituent (Jun 2005). In contrast, this percolation does not apply in Berber, in which the exact location of a tonal event is consequently more variable. Similarly, a language like Ambonese Malay (Maskikit-Essed and Gussenhoven 2016) may associate tones to phrasal constituents without percolation to specific words, let alone word heads (because these are not available). The same analysis may also be applicable to data from Australian languages, such as Dalabon and Mawng (Fletcher 2015, and Fletcher et al. 2016, respectively). Differentiating these two types of association allows us to distinguish between languages with very localized tonal events and languages in which the alignment of segments and tonal events is more diffuse. Logically, the same distinction should apply to languages that have lexical stress as well. Finally, this reanalysis of the relationship between phonological association and tonal alignment allows us to distinguish between the prosodic level relevant for tonal association and the level up to which percolation applies. These levels are likely to be tone- or language-specific.

A similar typological distinction may apply with respect to Grice's point about languages in which nuclear configurations are critical and those in which no special status is assigned to the final accent in a phrase. Overall, these additional dimensions of how tonal events function and how they behave phonologically and phonetically may provide more insightful categorizations than typologies based on phonetic detail (for discussions, see Ladd 2008; Arvaniti 2016; Prieto and Hualde 2016; chapter 1, section 1.5, this volume).

Using such distinctions in cross-linguistic work will allow us, as Grice also argues, to gain greater insight into un(der)described languages, without the imposition of a model that may best fit languages of a particular type, which simply happened to be the one that has been extensively studied. Studying prosodically diverse languages will further allow us to distinguish between marked and unmarked relationships between tune and text, such as whether the Spanish pattern described in Torreira and Grice (2018) is cross-linguistically frequent or represents a rare occurrence.

The Treatment of Phonetic Variability

This discussion had identified a need for new typology distinctions. Nevertheless, it is important to consider if and when additional categories are needed and whether simpler solutions that do not require additional theoretical machinery may be possible (Arvaniti, Ladd, and Mennen 2006). This is essential for avoiding overanalysis or unnecessarily complicating AM representations.

Grice's discussion of the Berber, Maltese, and Spanish data reflects this point, as it addresses an issue underlining much of AM research, namely AM's approach to variability. As I argue elsewhere as well (Arvaniti 2019; Lohfink, Katsika, and Arvaniti 2019), AM often treats variability as a major concern and has been unduly preoccupied by the search for tonal invariance. This search was originally motivated by the traditional autosegmental view that phonological association confers phonetic synchronization between associated autosegments (Goldsmith 1976). This view was supported by the interpretation of the findings of Arvaniti, Ladd, and Mennen (1998), a study originally motivated by the need to test competing views about the nature of intonational primitives as pitch movements or levels. In that study, we found that the rises of prenuclear pitch accents in Greek do not have stable characteristics, such as fixed slope or duration, as advocated by the Institute for Perception Research (IPO) approach ('t Hart, Collier, and Cohen 1990). Rather, in Greek, the position of both the onset and offset of the rise is stably synchronized with the onset and offset of the stressed syllable: the onset of the rise occurs slightly before the syllable onset; the peak of the rise occurs early into the vowel of the syllable following the stressed one. We concluded that this difference argues in favor of analytically treating the rise as the outcome of an L tone followed by an H tone, each with its own alignment: "The L and the H of the accent are *anchored* [emphasis added] to segmentally defined positions, and the duration and slope of the pitch movement are completely determined by the segmental composition of the accented word" (24). This finding gave rise to the notion of segmental anchoring, which was explored in detail in a series of papers by Ladd and colleagues (e.g., Ladd et al. 1999; Ladd, Mennen, and Schepman 2000; Ladd and Schepman 2003; Ladd et al. 2009) and later by many other researchers. Segmental anchoring soon became a main criterion for tonehood: pitch minima and maxima are treated as potential tonal targets reflecting underlying L and H tones, respectively. This conceptualization of the relation between phonological tones and phonetic realization led to a search for invariance in tonal alignment as a means of ascertaining whether a tonal target reflects an underlying tone. For example, Ladd and Schepman (2003) argue that the accent usually analyzed as H* in English (e.g., Pierrehumbert 1980) should be analyzed as (L + H)*, based on the finding that the dip before the rise to a peak is stable and thus must be the reflex of an underlying tone.

This search for invariance is questionable on two grounds. Statistically, it is the search to prove the null hypothesis, an overall inadvisable quest. Most importantly, variability is part and parcel of phonetics and thus unlikely to leave the realization

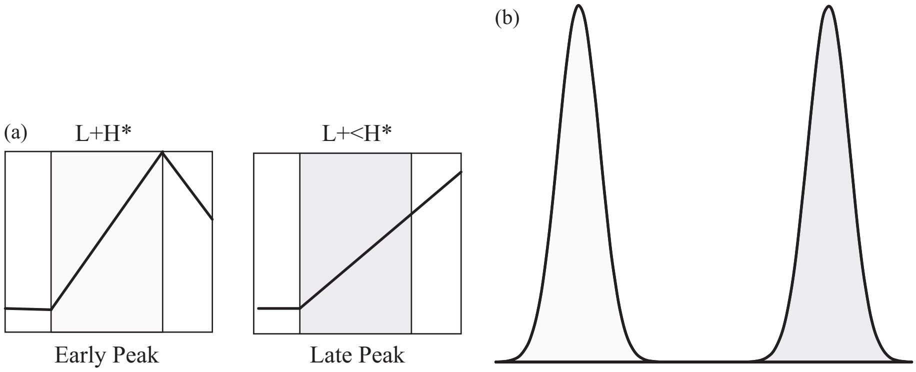


Figure 1r.1
In (a), schematic illustration of two Catalan accents, L + H* with early peak, i.e. co-occurring with the stressed syllable, and L + <H*, an accent with late peak, that is, occurring after the stressed syllable offset (after Prieto 2014); shaded rectangles represent the stressed syllable; in (b), idealized distributions of early and late peak alignment, as envisaged by the tonal anchoring approach.

of tonal categories unaffected. This is illustrated in figures 1r.1 and 1r.2. Figure 1r.1a shows schematic representations of an early peak and a late peak accent of Catalan (L + H* and L + <H*, respectively), as posited by Prieto (2014). Figure 1r.1b reflects how this phonetic difference must be distributed in a set of data if tonal anchoring and invariance hold: categories must have very narrow and completely distinct distributions, as idealized in the schematic density plots in figure 1r.1b.

However, as shown in figure 1r.2, such distributions are not attested in speech. This applies even for categories unequivocally considered distinct both phonologically and phonetically, such as the short-lag versus long-lag Voice Onset Time distinction in English. As shown in figure 1r.2a with data from Nakai and Scobbie (2016), even such categories show substantial overlap (for similar results, see also Piccinini and Arvaniti 2018). Thus, overlaps of similar magnitude should be expected for tonal categories. This is indeed the case, as illustrated in figure 1r.2b with data from Lohfink, Katsika, and Arvaniti (2019) regarding the H*, L + H* and H* + L of Greek.[1]

A possible reason for seeing such overlap as a problem is the typical visualizations of results coupled with the fact that the focus is on invariance and demonstrating that tonal categories are distinct along some dimension, such as alignment. Compare, for example, figure 1r.2c showing the box plots of the same data as in figure 1r.2b; the data show statistically significant differences by accent, but the box plots belie the extent of the overlap of the three accentual categories and, together with ideas about invariant alignment, may lead to unrealistic expectations about the realization of tonal categories.

An additional reason for the problem is the limited recognition within AM that the realization of tonal categories may rely not only on F0 but on additional cues as well. As Grice mentions, several studies show this connection, though this is typically seen as an effect of segments on tune or vice versa. Segmental material (or metrical structure, as Grice shows) can affect the choice or realization of the tune, while the tune may have repercussions for the realization of segments. An alternative perspective is to consider the latter type not as an adjustment, as Grice argues, but as the presence of redundant or secondary cues to a specific tonal contrast (much as a distinction between vowels of different length, or geminate and singleton consonants may also include a difference in phonetic parameters beyond duration; see Remijsen and Gilley 2008 on Dinka vowels; Local and Simpson 1999 on Malayalam geminates). More attention should be given to such additional cues and their relationship to F0, on the one hand, and to their role in the processing of intonation, on the other.

Phonetics, Phonology, and the Study of Intonation

The previous discussion and Grice's commentary reflect the uneasiness in AM (and other models of intonation, for that matter) regarding the relationship and division of labor between phonetics and phonology. This is not an issue that is by any means resolved in other domains. However, it remains particularly problematic in the study of intonation, which is often seen as a phenomenon on the edge, as being borderline linguistic but not quite fully so—hence, descriptions of intonation as a "half-tamed savage," as an element that may defy linguistic structure in some ways (see Gussenhoven 2004 for a discussion). Note that borderline phenomena are also found in segmental systems, as discussed in Scobbie (2007), but it is only in intonation that they are seen as the essence of the phenomenon itself.

In Arvaniti (2019) I argue against this view and in favor of a new approach, dubbed TINT (for *tame intonation*) that posits that intonation is not different from other components of a language's phonology. There are several corollaries to this approach,

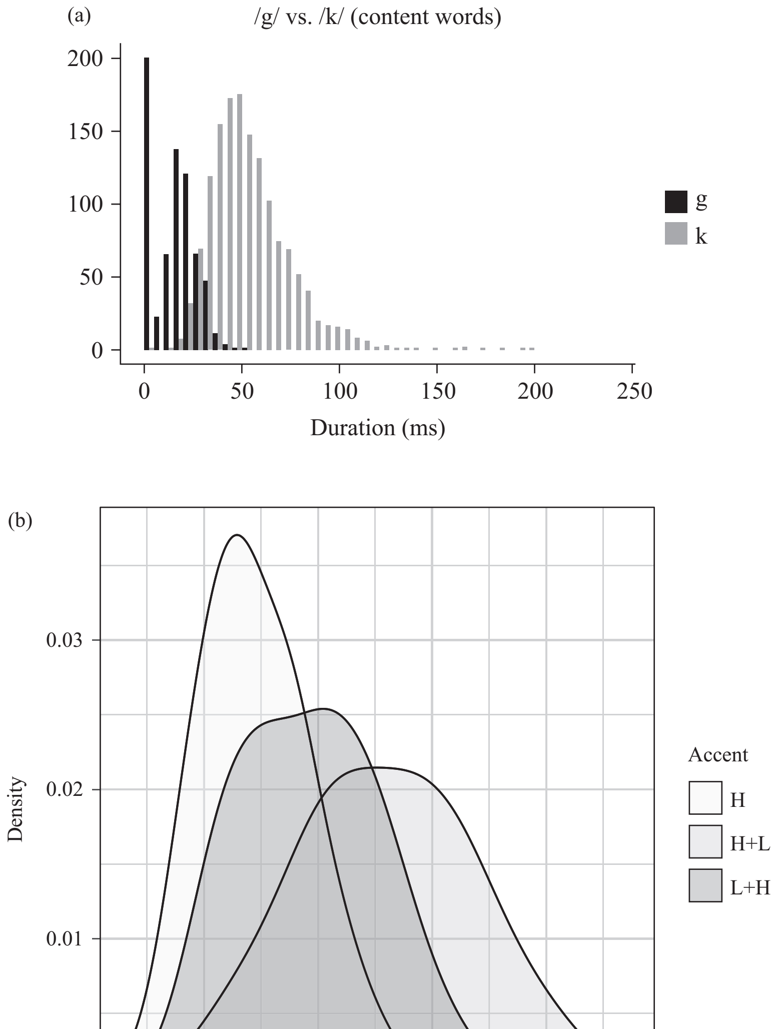


Figure 1r.2

In (a), VOT histograms for English [g] and [k] (short-lag and long-lag respectively) in word-initial position in content words (Nakai and Scobbie 2016); in (b), density plots of PC1 (first principal component in functional principal component analysis) for three Greek accents, H*, L + H*, and H* + L (from Lohfink et al. 2019); in (c) box-and-whiskers plots using the same data as in panel (b).

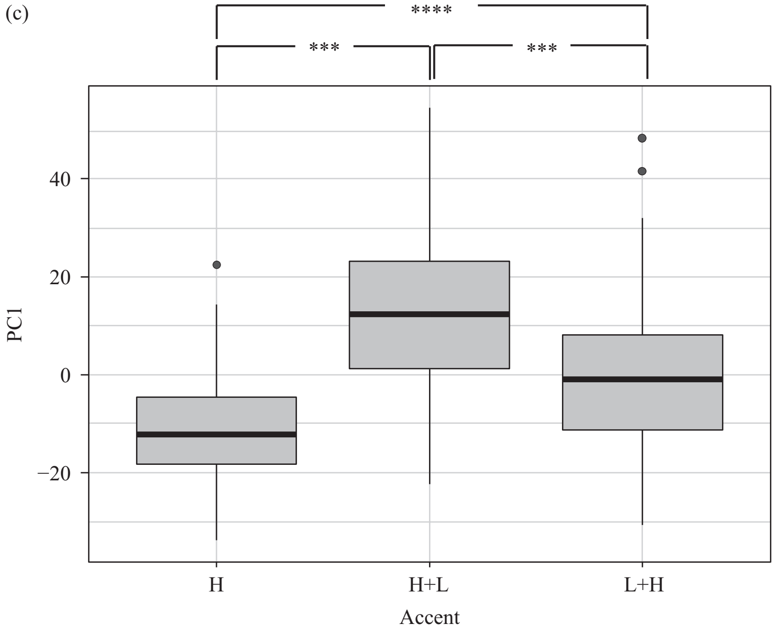


Figure 1r.2
(continued)

including a recognition that the realization of intonation is likely to present contextual, dialectal, stylistic, sociolinguistic, and speaker-specific variation and include phonetic dimensions beyond F0, which may be in trading relations with F0 itself (see Grice's commentary for some evidence). Furthermore, results such as those of Lohfink, Katsika, and Arvaniti (2019) indicate that this variation may be more pronounced for some tonal categories. For example, as shown in figure 1r.2b, the Greek H* accent is less variable in terms of PC1 in the functional principal component analysis, as compared to H* + L and L + H*; this can be interpreted as it being less variable in tonal scaling.

To be able to consider all of these facets of variation without losing sight of abstraction (which is essential for successful analysis, as argued by Arvaniti and Ladd 2009), it is important to adopt more flexible approaches, as argued by Grice, and newer techniques of examining data (cf. Lohfink, Katsika, and Arvaniti 2019). Finally, if the idea of tonal invariance is replaced or given lesser priority, AM needs to engage with pragmatic meaning, which should be an important criterion for positing phonological categories (Arvaniti 2007, 2016). Doing so should allow AM research to stop treating variation as problematic and will minimize overanalysis.

Note

1. The data come from functional principal component analysis of a corpus of 844 pitch contours elicited from thirteen speakers of Greek, following Gubian, Torreira, and Boves (2015). The data shown in figures 1r.2b and 1r.2c are from the first principal component (PC1), which accounts for 64.8 percent of the variability in the data and captures primarily scaling but also alignment differences between the accents (for details, see Lohfink, Katsika, and Arvaniti 2019).

References

Arvaniti, A. 2007. “On the Relationship between Phonology and Phonetics (or Why Phonetics Is Not Phonology).” *Proceedings of ICPhS XVI*, Saarbrueken, Germany, 19–24. http://www.icphs2007.de/.

Arvaniti, Amalia. 2016. “Analytical Decisions in Intonation Research and the Role of Representations: Lessons from Romani.” *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7 (1): 6. http://doi.org/10.5334/labphon.14.

Arvaniti, Amalia. 2019. “Crosslinguistic Variation, Phonetic Variability, and the Formation of Categories in Intonation.” *Proceedings of the Nineteenth International Congress of Phonetic Sciences,* Melbourne, Australia, 2019, edited by Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren, 1–6. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Arvaniti, Amalia, and D. Robert Ladd. 2009. “Greek wh-Questions and the Phonology of Intonation.” *Phonology* 26:46–63.

Arvaniti, Amalia, D. Robert Ladd, and Ineke Mennen. 1998. “Stability of Tonal Alignment: The Case of Greek Prenuclear Accents.” *Journal of Phonetics* 26:3–25.

Arvaniti, Amalia, D. Robert Ladd, and Ineke Mennen. 2006. “Tonal Association and Tonal Alignment: Evidence from Greek Polar Questions and Contrastive Statements.” *Language and Speech* 49:421–450.

Fletcher, Janet. 2015. “Intonation and Prosody in Dalabon.” In *Prosodic Typology II: The Phonology of Intonation and Phrasing*, edited by Sun-Ah Jun, 252–272. Oxford: Oxford University Press.

Fletcher, Janet, Hywel Stoakes, Ruth Singer, and Debbie Loakes. 2016. “Intonational Correlates of Subject and Object Realization in Mawng (Australian).” In *Proceedings of Speech Prosody 2016,* edited by Jonathan Barnes, Alejna Brugos, Stefanie Shattuck-Hufnagel, and Nanette Veilleux, 188–192. http://www.isca-speech.org/archive/SpeechProsody_2016/.

Goldsmith, John A. 1976. “Autosegmental Phonology.” PhD diss., MIT.

Gubian, Michele, Francisco Torreira, and Lou Boves. 2015. “Using Functional Data Analysis for Investigating Multidimensional Dynamic Phonetic Contrasts.” *Journal of Phonetics* 49:16–40.

Gussenhoven, Carlos. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press. Jun, Sun-Ah. 2005. “Korean Intonational Phonology and Prosodic Transcription.” In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by Sun-Ah Jun, 201–229. Oxford: Oxford University Press.

Ladd, D. Robert. 2008. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Ladd, D. Robert, Dan Faulkner, Hanneke Faulkner, and Astrid Schepman. 1999. “Constant ‘Segmental Anchoring’ of F0 Movements under Changes in Speech Rate.” *Journal of the Acoustical Society of America* 106 (3): 1543–1554.

Ladd, D. Robert, Ineke Mennen, and Astrid Schepman. 2000. “Phonological Conditioning of Peak Alignment in Rising Pitch Accents in Dutch.” *Journal of the Acoustical Society of America* 107:2685–2696.

Ladd, D. Robert, and Astrid Schepman. 2003. “‘Sagging Transitions’ between High Pitch Accents in English: Experimental Evidence.” *Journal of Phonetics* 31:81–112.

Ladd, D. Robert, Astrid Schepman, Laurence White, Louise May Quarmby, and Rebekah Stackhouse. 2009. "Structural and Dialectal Effects on Pitch Peak Alignment in Two Varieties of British English." *Journal of Phonetics* 37 (2): 145–161.

Local, John, and Adrian P. Simpson. 1999. "Phonetic Implementation of Geminates in Malayalam Nouns." *Proceedings of International Congress of Phonetic Sciences 1999*, San Francisco, edited by John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville, and Ashlee C. Bailey, 595–598; ICPhS Archive. http://www.internationalphoneticassociation.org/icphs/icphs1999.

Lohfink, Georg, Argyro Katsika, and Amalia Arvaniti. 2019. "Variability and Category Overlap in the Realization of Intonation." In *Proceedings of the Nineteenth International Congress of Phonetic Sciences*, Melbourne, Australia 2019, edited by Sasha Calhoun, Paola Escudero, Marija Tabain, and Paul Warren, 701–705. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Maskikit-Essed, Raechel, and Carlos Gussenhoven. 2016. "No Stress, No Pitch Accent, No Prosodic Focus: The Case of Ambonese Malay." *Phonology* 33:353–389.

Nakai, Satsuki, and James M. Scobbie. 2016. "The VOT Category Boundary in Word-Initial Stops: Counter-evidence against Rate Normalization in English Spontaneous Speech." *Laboratory Phonology* 7 (1): 13. http://doi.org/10.5334/labphon.49.

Piccinini, A., and A. Arvaniti. 2018. "Dominance, Mode, and Individual Variation in Bilingual Speech Production and Perception." *Linguistic Approaches to Bilingualism* 9 (4–5): 628–658. https://doi.org/10.1075/lab.17027.pic.

Pierrehumbert, Janet B. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pierrehumbert, Janet, and Mary Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Prieto, Pilar. 2014. "The Intonational Phonology of Catalan." In *Prosodic Typology II: The Phonology of Intonation and Phrasing*, edited by Sun-Ah Jun, 43–80. Oxford: Oxford University Press.

Prieto, Pilar, and José Ignacio Hualde. 2016. "Towards an International Prosodic Alphabet (IPrA)." *Laboratory Phonology* 7 (1): 5. http://doi.org/10.5334/labphon.11.

Remijsen, Bert, and Leoma Gilley. 2008. "Why Are Three-Level Vowel Length Systems Rare? Insights from Dinka (Luanyjang Dialect)." *Journal of Phonetics* 36 (2): 318–344.

Scobbie, James, M. 2007. "Interface and Overlap in Phonetics and Phonology." In *The Oxford Handbook of Linguistic Interfaces*, edited by Gillian Ramchand and Charles Reiss. https://doi.org/10.1093/oxfordhb/9780199247455.013.0002.

't Hart, Johan, René Collier, and Antonie Cohen. 1990. *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Torreira, Francisco, and Martine Grice. 2018. "Melodic Constructions in Spanish: Metrical Structure Determines the Association Properties of Intonational Tones." *Journal of the International Phonetic Association* 48 (1): 9–32.

2 Modeling Danish Intonation

Nina Grønnum

2.1 Introduction

Danish belongs to the Nordic (North Germanic) group of Germanic languages. It is the national language of Denmark, and according to Lützen (2010), it is the mother tongue of 92 to 93 percent of Denmark's 5.82 million inhabitants (as of 2020). It is written with the Latin alphabet, extended with *æ*, *ø*, and *å*. Grønnum (2000)[1] is a very brief, and Grønnum (2007) a more comprehensive, modern introduction to its phonetics. The phonology is exhaustively described in Basbøll (2005).

There are three prosodic components in the Danish sound system: stress, stød, and intonation. If vowel length is considered a syllable prosody, it will make a fourth. The phonology of stress and stød is accounted for in Basbøll (2005), and Grønnum, Vazquez-Larruscaín, and Basbøll (2013) contains a more recent discussion specifically about the abstract representation of stød.

This chapter is about intonation only. It is a summary of acoustic analyses and a few perceptual experiments carried out from the mid-1970s through the 1990s (Thorsen 1978 through 1988; Grønnum 1990 through 1998). It presents a model of the acoustic output from Copenhagen Danish speakers' production in distinct, formal speech, of various types of utterances, as well as short texts. Comparison with Central Swedish will highlight the characteristic properties of Danish intonation.

2.2 Previous Descriptions

The first linguist to concern himself with Danish intonation was the brilliant J. P. Høysgaard (1747, 1769). According to him, stressed syllables are associated with high tone and unstressed syllables with low tone. He denies intonation any real linguistic status, because words can be comprehended even when whispered. Tone helps only to make spoken language clearly audible.

In the literature of the nineteenth and first half of the twentieth centuries, there is likewise an almost general consensus that stressed syllables have high pitch (Jensen 1833; Sneedorff-Birch 1835; Jerndorff 1897, 1911; Forchhammer 1903; Oxenvad 1937; Arnholtz 1939–1940, 1942, 1954). However, Andersen (1954) notes that in North Zealand and in Copenhagen, stressed syllables are spoken on a lower pitch. To Arnholtz, this is characteristic of the vulgar speech of the lower classes in Copenhagen, where a loathsome change in the stress~pitch relation is taking place: stressed syllables with low or falling pitch—traditionally called *omlæggertone* (approximately "reverse tone"). As recently as 1975, Brink and Lund repeat Arnholtz's observation (albeit in less condemning terms) and state (608–610) that Low Copenhagen often has a falling pitch movement where High Copenhagen has a rising pitch, and vice versa, Low Copenhagen rises where High Copenhagen falls. Jerndorff (1897) observes that the degree of relative prominence on each stressed syllable correlates directly with its pitch height. The more weight a word is given, the higher the pitch of its stressed syllable. But he added that highlighting of a stressed syllable can be achieved by lowering the preceding unstressed syllable as well.

There are also scattered remarks about sentence modality and tone. Statements are falling and/or end low, and questions are rising and/or end high—or at least they are not falling (Bloch 1805; Mikkelsen 1894; Jerndorff 1897; Forchhammer 1903; Oxenvad 1937, Hjelmslev 1954). In Hjelmslev's (1951) terminology, there are two modulations: one falling and one nonfalling. He notes specifically that these modulations cannot be partitioned into smaller structural domains. Jerndorff (1911) takes issue with the rule for declamation (in schools, lecture halls, churches, and on the stage) that questions must end in a final rise through the unstressed syllables. Other languages may do so, but Danish does not, and he ascribes the role of *carrier of intonation* to the stressed syllables only. Arnholtz (1954) likewise states that in the Danish standard language, it seems that the stressed syllables carry the speech melody, the unstressed syllables being subordinate to the pitch pattern established by the stressed ones.

In unpublished lecture notes from seminars in 1946 and 1947 about Danish speech rhythm and melody, Eli Fischer-Jørgensen makes the same assertion about stressed syllables being the carriers of the speech melody. She discusses at some length the possible criteria for delimiting Danish utterances into stress groups, particularly the problem of assigning the boundaries between them.

Until the 1970s, the only comprehensive accounts of Danish prosody were Otto Jespersen's *Fonetik* (1897–1899, chapters 25 and 26), excerpts of which are found in his *Modersmålets Fonetik* (1934, chapters 13–16), and Alf Bo's monograph *Tonegangen i Dansk Rigsmål* (1933). Jespersen (1897–1899) talks about the law of termination: a low tone or fall in pitch is employed when we are through and want to finish, while a rise to a higher pitch indicates the nonfinished, nonterminal. Bo establishes rules of intonation mainly by impressionistic observation of his own speech. Questions have rising intonation, but the magnitude of the rise in questions varies. Questions with no lexical or syntactic information about their interrogative status are the most rising; questions with word order inversion or interrogative particle less so and may approach declaratives.

Thus, the general picture that emerges from the Danish literature is one where stressed syllables are normally associated with relatively higher pitch than unstressed ones; intonation is carried by the stressed syllables; statements are associated with falling intonation and questions with rising intonation. It is not quite clear whether rises and falls are situated at the end of the utterance or whether they characterize the entire succession of stressed syllables. There are hints that intonation contour trends may be global, especially in some of the transcriptions, but there are also rather clear statements to the effect that only the last stressed syllable is affected.

2.3 Preliminaries

The ultimate goal of prosodic theory or, in this case, intonation theory, would be a representation that in a psychologically real way faithfully models every aspect of speakers' and listeners' processing of intonation, including syllable and word tones, in spoken language in general. That is obviously not yet within reach. Settling for a more modest goal, intonation can be modeled from one of several specific vantage points:

as speakers' production or speakers' perception; or as input to synthetic speech, or as patterns for speech recognition—procedures that do not necessarily parallel human processes. Furthermore, speaker and listener processing may be modeled at different levels of abstraction. The first approximation, described in figure 2.10, models two components and their interaction in the production of short utterances in the formal speech of speakers of Copenhagen Danish, leaving out phenomena that typically are not under the speaker's voluntary control, that is, *microprosody*. The more elaborate edition of the model, in figure 2.13, expands the structural domain to longer and syntactically more complex utterances and short texts.

Ideally, a model of the production of Danish intonation is a representation that mediates between a stage upstream in the human production of linguistic utterances, where all the morphosyntactic and lexical information has been supplied and stress assigned to the appropriate syllables, and a stage downstream where significant intonational information is inserted. From this latter stage the phonetic fundamental frequency implementation can be derived by rule and converted to neural commands to drive the peripheral physiological production system or, alternatively, converted to commands to drive a speech synthesizer. Over the last couple of decades, it has become customary to call a representation that in this manner mediates between a less and a more concrete stage in the production of intonation *phonological*. Ladd (2008) contains a lucid discussion of the pros and cons of such a concept. However, if you want to avoid the obvious—but obviously false—analogy to segmental phonology and its sequentially ordered and minimally contrastive units, then *symbolic representation* is perhaps a better concept.

2.4 Danish Intonation

It is customary to let "intonation" designate the linguistically relevant but nonlexical aspects of the *course of fundamental frequency* (or its perceptual correlate, the *pitch course*) through spoken utterances (see, e.g., Ladd 1996). Nonlexical leaves out syllable tones and word accents, but it also excludes the contribution to fundamental frequency (henceforth F_0) that comes from stress. However, the stress~F_0 relation is a salient property of Danish speech melody. Therefore, as long as the subject is Danish, which has neither syllable tones nor tonal word accents, let intonation designate the linguistically relevant aspects of F_0 in spoken utterances, including F_0 patterns associated with stress. "Linguistically relevant" excludes microprosodic effects, that is, the contribution from intrinsic F_0 properties of segments and perturbations at segment boundaries. I take the core linguistic function of intonation to be to signal (i) prominence relations, (ii) prosodic boundaries between phrases and utterances, and (iii) utterance modality.

2.4.1 Empirical Investigations

Empirical studies of Danish intonation were long in coming, even well after the appearance of suitable instrumentation for analysis. Intonation was considered difficult to approach, for reasons stated by Poul Andersen (1954, 310) [in my near verbatim translation]: "To give a general description of tonal, stress, and speech rate variations in the sentence is made difficult by the fact that these phenomena vary with the speaker's mood and emotional state, and it is very difficult to decide if we are dealing with, e.g., a neutral question or a surprised question, and, strictly speaking, there really is no such thing as a neutral intonation, since all speech is connected with some emotional state." In other words, F_0 codes several types of information simultaneously—for example, the linguistic function, the affective state of the speaker, the speaker's age, gender, personal identity, and membership of regional and social groups. The difficulty inherent in teasing out the contribution from each of these various factors likely appeared prohibitive.

By the mid-1970s, however, there was a growing demand for descriptions of intonation that could be applied in the teaching of Danish as a foreign language and in teaching the hearing impaired. In spite of Andersen's apprehensions, I set out to discover the linguistically relevant properties in the "tonal variations in the sentence" and provide a description of "neutral intonation." Andersen is right that, strictly speaking, there is no such thing as a neutral utterance or neutral intonation. Every utterance is produced in some context, and the context is one factor in the manner of its production. But we can operationally define two helpful concepts: *unmarked* intonation and *typical* intonation. By definition, subject-verb (SV) declarative sentences are syntactically unmarked in Danish. By convention, the intonation contour that accompanies such an SV declarative when it is used conventionally is unmarked. Typical intonation, on the other hand, is one that accompanies any given sentence type when used conventionally. Thus, for instance, the intonation associated with a conventionally used yes/no question will be different from the unmarked intonation of conventional declaratives; hence it is marked, but that is typical.

2.4.2 Methodology

The only way to approach the relation between stress and F_0 in typical intonation contours, keeping everything else constant and similarly, to lay bare typical intonation contours, was to have speakers read material designed for the purpose—in this case, short sentences embedded in small dialogues, some of which were built around nonce words with systematically different word stress locations. Details can be found in Thorsen (1978), and suffice it here to say that in this first experiment, as well as every succeeding one, there were always at least four speakers and at least six repetitions of every target utterance by each speaker in randomized scripted material. Note that everything I say about Danish as well as Swedish intonation in the following pertains to utterances conventionally used, recorded in a formal and distinct style of speech.

2.4.3 Paucity of Prosodic Components in Danish

Almost every significant fact about Danish intonation can be deduced from the declarative utterance in figure 2.1, with its unmarked and typical declarative intonation

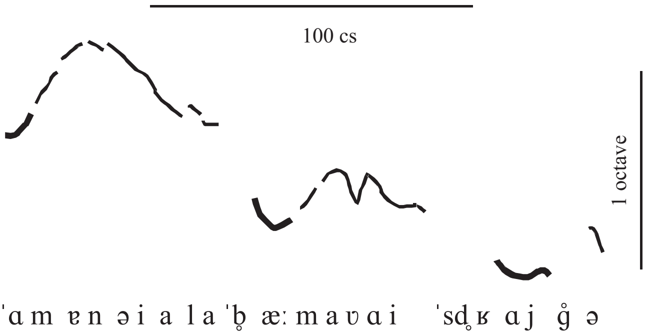


Figure 2.1
Fundamental frequency tracing of a female speaker's utterance: *Ammerne i Alabama var i strejke*. The stressed vowels are in heavier lines.

contour: [ˈɑmɐnə i alaˈb̥æːma ʋɑ i ˈsd̥ʁɑjg̊ə] *Ammerne i Alabama var i strejke* (The wet-nurses in Alabama were on strike), composed so as to minimize confounding influence from differences in intrinsic vowel F_0 and perturbations at segment boundaries. (This is merely a sample utterance for illustration, backed up by the results of empirical investigations.)

- If the course of F_0 is any indication, the six words in this utterance form three prosodic groups: {Ammerne i Ala}{bama var i}{strejke}. Each is characterized by a rise in F_0 from the stressed to the first posttonic syllable, a rise whose magnitude decreases from start to finish, followed by a more or less extensive fall through succeeding posttonics, if any. (The small dip halfway through the second pattern is due to the intrinsically lower F_0 of the [ʋ].) The extent of the F_0 fall varies in accordance with the number of posttonic syllables, and with only one posttonic syllable (as in *strejke*), the fall is absent.
- There is no special and elaborate F_0 movement at the end of the utterance.
- There is no pronounced fall to low at the very end.
- The stressed diphthong of *strejke* is not visibly or perceptibly lengthened.

In similar circumstances, a Central Swedish utterance—in addition to the tonal patterns associated with the word accents—would have exhibited a prominent final rising sentence accent, a clear final low boundary tone, and final lengthening—all of which are absent here.

2.4.4 The Prosodic Stress Group and Its F_0 Pattern

F_0 is the most explicit among the acoustic cues to stress. The onset of any stressed vowel coincides with the onset of a recurrent melodic pattern that extends over all succeeding unstressed syllables, irrespective of their morphological or syntactic affiliation, cutting across word boundaries, until the onset of the next stressed vowel. Such a sequence of syllables forms a prosodic stress group, with an associated F_0 pattern. This is the stress group that Fischer-Jørgensen intuited in 1946–1947, but whose boundaries she hesitated to determine. Its F_0 pattern, with the low stressed syllable relative to the first posttonic, is the one that Arnholtz (1939–1940) found vulgar and loathsome. It is now the norm in Copenhagen Danish and no longer confined to the "lower classes," as Fischer-Jørgensen (1983) also noted in her acoustic analysis of a Copenhagen speaker. The pattern is subject to truncation/extension (rather than compression/expansion), as evidenced in figure 2.1. The shorter or longer the stress group, the less or more extensive is its F_0 pattern. It is also sensitive to location: rises (and falls) are more extensive early in the utterance than later, and mostly also slightly more extensive on less declining intonation contours, everything else being equal.

2.4.5 Equal and Unequal Prominence

In Standard Danish, in pragmatically and emotionally neutral speech, no one stressed syllable is more prominent than any other. Such prominence is not evident in acoustic registrations (in duration and/or the F_0 contour) nor is it present auditorily. There is nothing incomplete about such prosodically neutral utterances, and although they may not be very frequent in spontaneous speech, they certainly do occur, they are not unnatural, and they are very easy to elicit from speakers in a reading situation. This is a completely uncontroversial fact among Danish scholars (Andersen 1954; Basbøll 2005; Heger 1981; Rischel 1980, 1983).

2.4.5.1 No default sentence accent Merely looking at the tracing in figure 2.1, you might think that *Ammerne*, with its more ample F_0 pattern, sounded more prominent than *strejke*. That is not the case. The three stressed words are equally prominent perceptually. Nor do I expect anyone to read a final sentence accent into the utterance in figure 2.1. A comparison with Swedish is illustrative.

Stockholm and Copenhagen speakers recorded the same utterance in isolation, and in two different dialogue contexts: *Torbens syster heter Kamma/Torbens søster hedder Kamma* (Torben's sister is called Kamma). One context inquired about Torben's sister's name, to elicit focus on *Kamma*; the other asked who had a sister called *Kamma*, to elicit focus on *Torbens* and reduced prominence (deaccentuation) on *Kamma*. The six F_0 contours in figure 2.2 depict the last word, *Kamma*, in one recording of each of these six utterances. They are clearly different among themselves in Swedish. In the isolated utterance (left), *Kamma* exhibits a word accent II fall from high, a default sentence accent rise, and a (not very explicit) boundary low, as specified by Bruce (1977). When focused (middle), the sentence accent rise is higher and the fall to low deeper. When the focus is earlier in the utterance (right), the sentence accent rise is absent. The three words in the Danish utterances exhibit nothing comparable and are hardly distinguishable.

2.4.5.2 Focus by reduction The isolated Danish utterance, *Torbens søster hedder Kamma*, and the answer that focuses on *Kamma*, from which the tracings in figure 2.2 are drawn, are not distinguishable perceptually when played back out of context, which is not surprising when you look at the top and middle panels in figure 2.3. But that is specific to final words where there is no succeeding stress group to reduce, which is how focus is achieved in this speech style. Thus, the answer to whose sister is called *Kamma*, which focuses on *Torbens*, is clearly distinguishable from the isolated utterance, due not so much to increased F_0 on *Tor-*, but rather to the reduced prominence on succeeding *søster*, via deletion of the F_0 rise from *søs-* to *-ter* (bottom panel in figure 2.3).

2.4.5.3 Emphasis for contrast by reduction With a greater degree of relative prominence, the prominent word itself may be affected, that is, be boosted by a somewhat

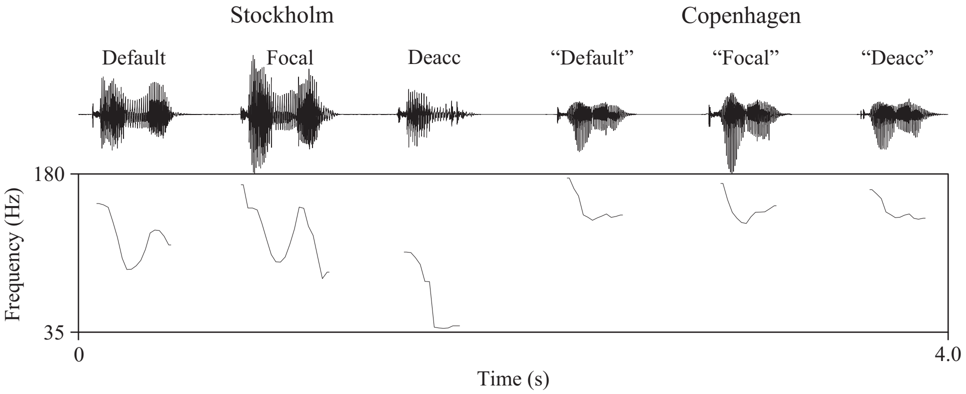


Figure 2.2
Fundamental frequency tracings of the last word, *Kamma*, in three short utterances in Stockholm Swedish (male speaker, left) and Copenhagen Danish (female speaker, right).

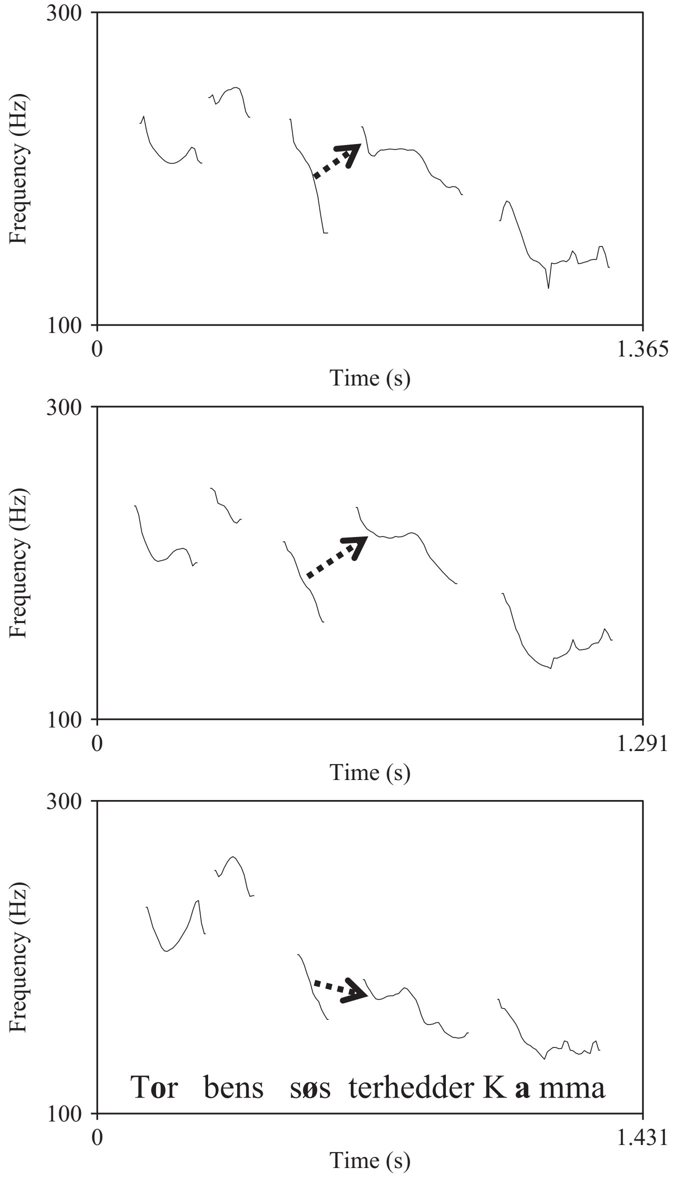


Figure 2.3
Three versions of *Torbens søster hedder Kamma* (female speaker): in isolation (top), with focus on *Kamma* (middle), and on *Torbens* (bottom).

increased F_0 on the stressed syllable, as well as a higher rise to the posttonic syllable. But equally significant is the reduction on preceding as well as succeeding stress groups, as illustrated in figure 2.4. It depicts an utterance, *Der går mange busser fra Tiflis* (There are many buses out of Tiflis), in isolation and as a response to a question that inquired whether there were in fact buses out of Tiflis or if one had to go by train. The F_0 rises on *mange* and *Tiflis* are deleted, and the rise from *bus-* is higher in the emphasized version on the right. Jerndorff's (1897) observation, that the degree of relative prominence on

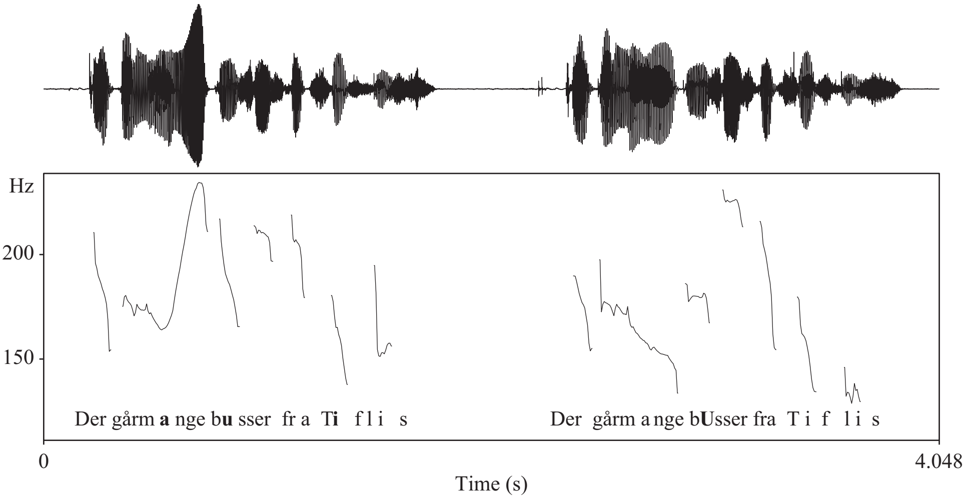


Figure 2.4
Fundamental frequency tracings of two versions by a female speaker of *Der går mange busser fra Tiflis* uttered in isolation (left) and with emphasis for contrast on *busser* (right).

a stressed syllable need not be achieved by increasing its pitch but can be accomplished also by lowering the preceding unstressed syllable, was not wide of the mark.

2.4.6 No Final Boundary Tone

The absence of a low, final boundary tone in Copenhagen is very apparent, also in comparison with Danish as spoken on Bornholm, a Danish island just south of Sweden in the Baltic Sea. Figure 2.5 depicts F_0 tracings of a long declarative utterance: *Kofoed og Thorsen skal med rutebilen fra Gudhjem til Snogebæk klokken fire på tirsdag* (Bornholm, top panel) and *Kofoed og Thorsen skal med bussen fra Fuglebjerg til Sorø klokken fire på tirsdag* (Copenhagen, bottom panel) (Kofoed and Thorsen are taking the bus from . . . to . . . at four o'clock on Tuesday). For ease of comparison, the frequency scale spans the same interval in both panels.

2.4.7 No Final Lengthening

Final syllables generally (but not invariably) exhibit a modest increased duration in Copenhagen Danish, everything else being equal. I used to think of that as final lengthening. However, when I compared it with Central Swedish, I had occasion to reconsider. Figure 2.6 depicts *Kamma* from initial and final position, respectively, in Swedish *Kamma kommer från Svaneke* and *Torbens syster heter Kamma* (left) and Danish *Kamma stammer fra Næstved* and *Torbens søster hedder Kamma* (right) (Kamma comes from Svaneke/Næstved; Torben's sister is called Kamma). The Swedish *Kamma* is 33 percent longer in final than in initial position, against a meager and nonperceptible 7 percent lengthening in Copenhagen. No doubt final lengthening is an independent factor in Swedish, but in Danish, it may be merely the result of a general articulatory relaxation toward the end of phrases and utterances, and not an autonomous, independently speaker-controlled feature. I would argue similarly about the increased phrase final duration that Tøndering (2008) found in the nonscripted monologues in the Danish Phonetically Annotated Spontaneous Speech corpus (DanPASS; see Grønnum 2009). It is unquestionably an acoustic reality, although *not* ubiquitous, but it is nowhere near

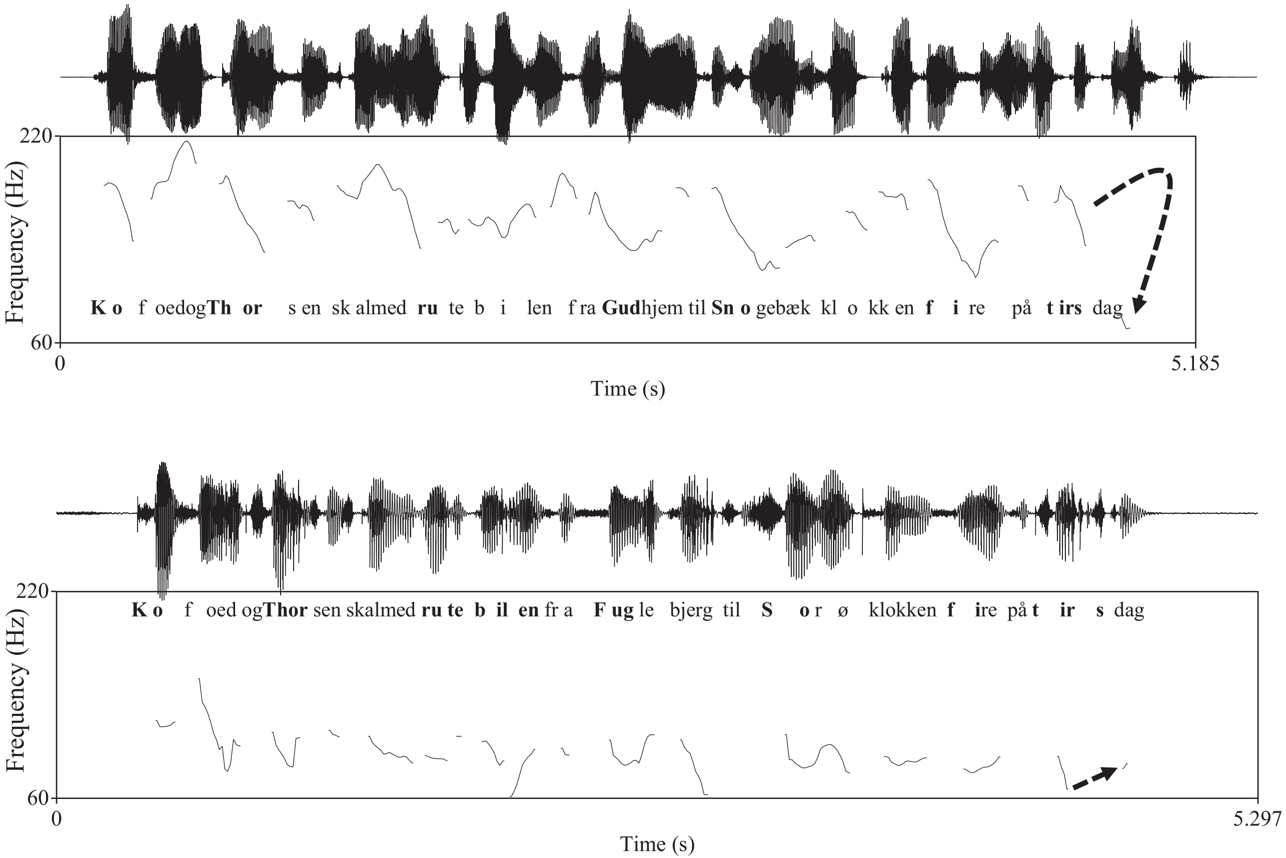


Figure 2.5
Fundamental frequency tracings of a long, terminal declarative utterance by male speakers in Bornholm (top) and Copenhagen (bottom).

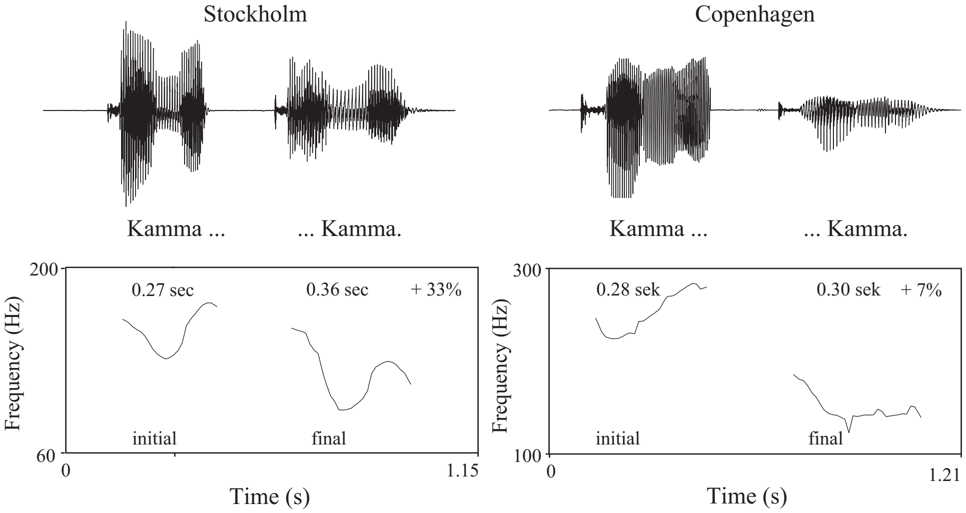


Figure 2.6
Fundamental frequency contours of the word *Kamma*, extracted from utterance initial and final position, respectively, in Swedish (left trace, male speaker) and Danish (right trace, female speaker) utterances.

as perceptually salient as in Central Swedish. In other words, final lengthening is not a component in the abstract, symbolic representation of Copenhagen Danish prosody.

2.4.8 No Final Rise or "High" in Questions

Compare Danish with Swedish once again. Among my recordings are the following utterances, produced in isolation: *Kamma kommer från Svaneke* (Stockholm; Kamma comes from Svaneke) and *Anders og Kamma skal til Fakse* (Copenhagen; Anders and Kamma are going to Fakse). I have small dialogues as well: *Hur långt är det från Sandvik till Svaneke?–Till Svaneke?–Det är tre mil* (Stockholm; How far is it from Sandvik to Svaneke?–To Svaneke?–It is three miles), and *Hvor langt er der fra Næstved til Fakse?–Til Fakse?–Der er cirka 30 kilometer.* (Copenhagen; How far is it from Næstved to Fakse?–To Fakse?–It is about 30 kilometers).

Figure 2.7 depicts F_0 tracings of the last word in the Stockholm and Copenhagen renderings, from the declarative (leftmost of three), the *wh*-question (middle), and the question that is not syntactically marked as such (an intonation question; rightmost of three). As expected, Stockholm exhibits the word accent I low stressed syllable, the sentence accent rise, and the boundary low, where the distinguishing feature is the magnitude of the sentence accent rise, which is somewhat higher in the *wh*-question and very much higher in the intonation question than in the statement. It is not that statements fall or end low in the Swedish utterances and questions rise or end high. Statements and questions alike end with a low boundary tone, but the preceding rise and sentence accent peak are significantly higher in questions, and much more so when there is nothing but the intonation to carry the interrogative modality. In the three Danish words, it is not apparent that the rise itself is significantly higher from *Fak-* to *-se* in either of the two questions, but the whole word is lifted way up in the range in the intonation question. So perhaps there is some truth to an assertion that questions—at least when devoid of lexical or syntactic cues to their modality—end

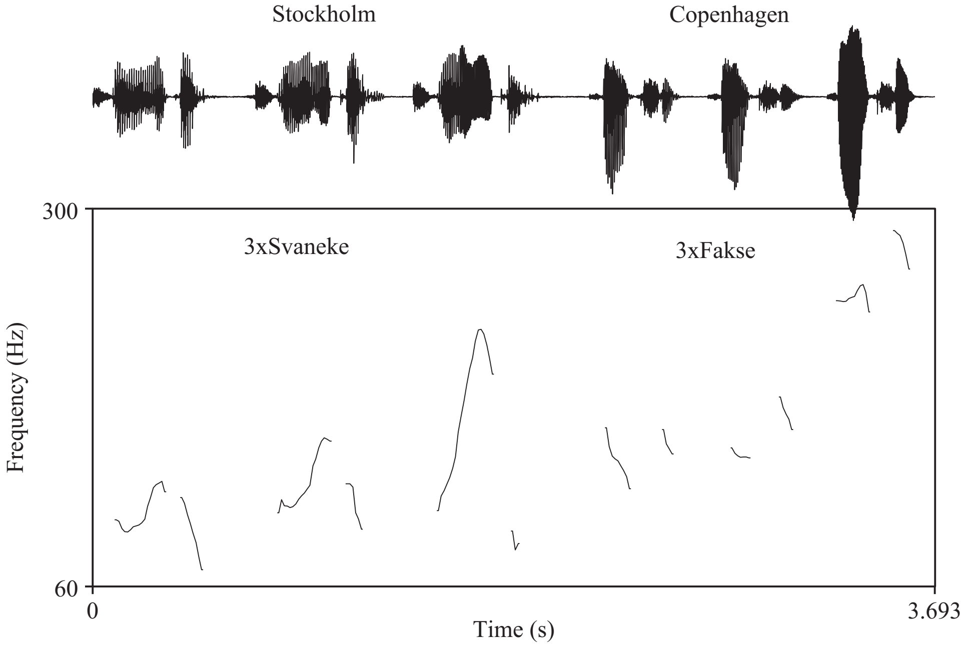


Figure 2.7
Fundamental frequency tracings of the last word (*Svaneke* and *Fakse*, respectively) in a statement and two questions in Stockholm (male speaker, left) and Copenhagen (female speaker, right).

high. Such a claim, however, misses the point. The pertinent feature becomes apparent when we look at a complete sentence, as in figure 2.8.

Compare the same sentence, *Der går mange busser fra Tiflis* (There are many buses out of Tiflis), uttered as a terminal declarative in isolation and as a question. The question was the third and last utterance in this little exchange: *Hvornår skal vi afsted? Hvad står der i køreplanen? Der går mange busser fra Tiflis?* (When are we leaving? What does the timetable say? There are many buses out of Tiflis?). It depends on the intonation to be perceived as a request for confirmation. What distinguishes this, the most marked of intonation contours, from the unmarked, declarative intonation, is not the final F_0 movement but rather the overall nondeclining contour, suggested by the broken lines. Because the contour is not actually rising, *Tiflis* in the intonation question on the right is not higher than the preceding *busser*. In this light, the rightmost *Fakse* in figure 2.7 should be seen not as high per se but as the termination of a (virtual) nondeclining contour.

There is more to be said about intonation and modality, but that is more easily done after introduction of the initial version of the model presented in section 2.4.10 and figure 2.10.

2.4.9 The Stressed Syllables Carry the Intonation Contour

Jerndorff's (1911) and Fischer-Jørgensen's (1946–1947) impression that the stressed syllables carry the intonation is entirely justified. Even more to the point is Arnholtz's (1954) added observation: that the unstressed syllables are subordinate to the pattern established by the stressed ones. It is evident in figure 2.1 that F_0 patterns are sensitive to location and diminish through the utterance. Figure 2.9 illustrates how the scaling of

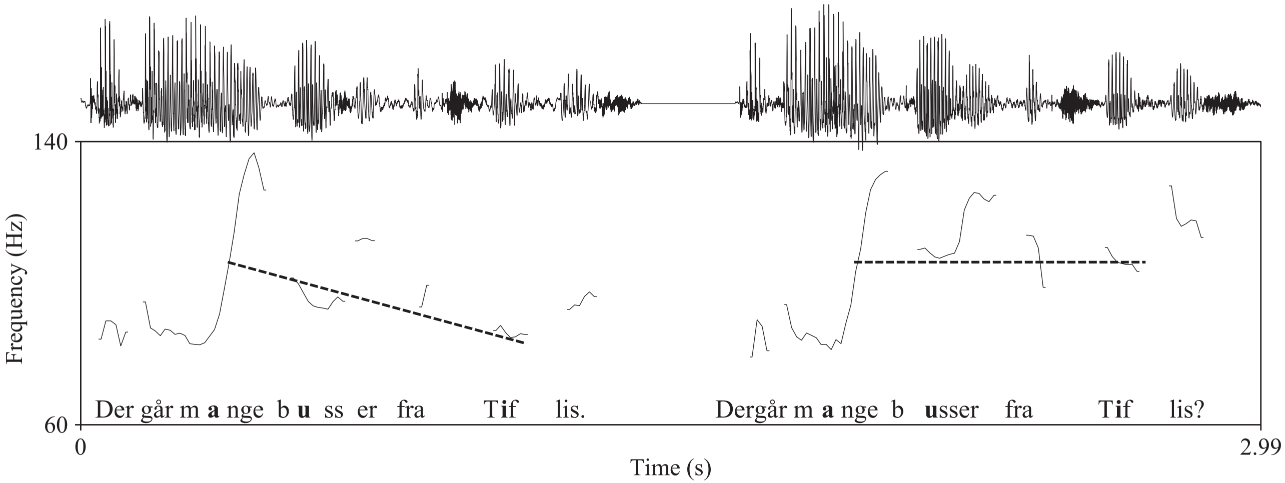


Figure 2.8
Fundamental frequency tracings of the same sentence produced as a terminal declarative (left) and a question (right), male speaker. Dotted lines are drawn through the three stressed vowels.

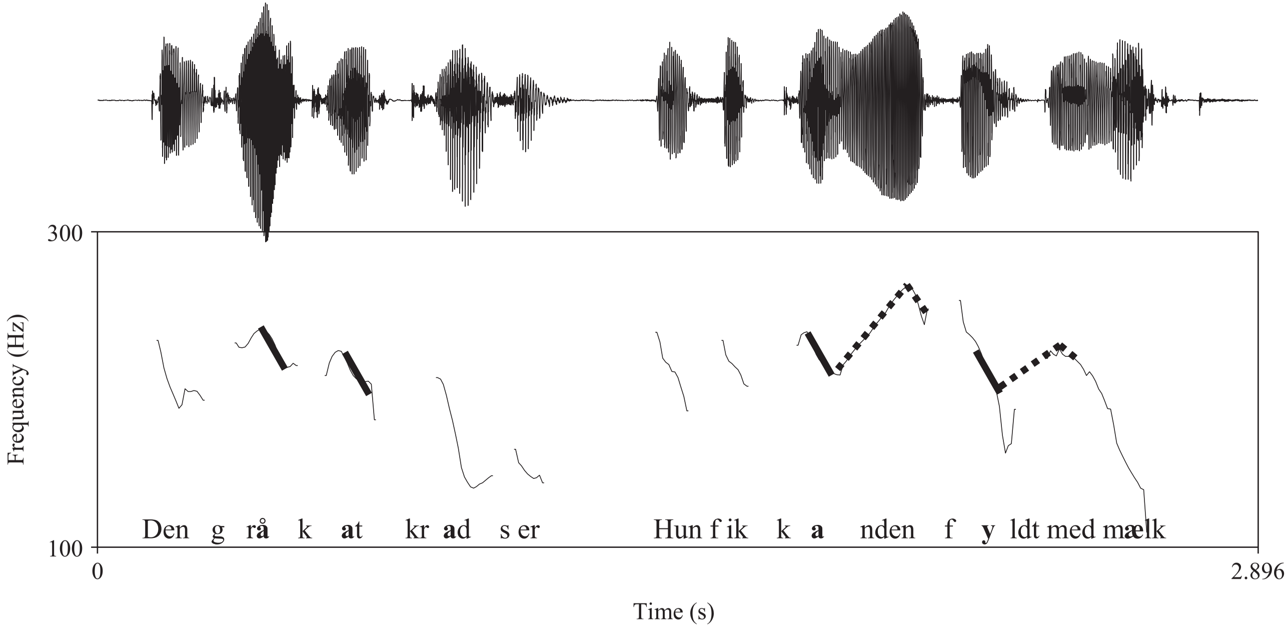


Figure 2.9
Fundamental frequency tracings of *Den grå kat kradser* (left) and *Hun fik kanden fyldt med mælk* (right), female speaker.

the stressed syllables is unaffected by the presence, or not, of unstressed syllables in the stress group. On the left is an utterance, *Den grå kat kradser* (The gray cat scratches), where the first two stressed syllables are without posttonics. On the right is *Hun fik kanden fyldt med mælk* (She had the jug filled with milk). *Grå* and *kan*- appear on approximately the same level, as do *kat* and *fyldt*, whether or not they are succeeded by rises to a posttonic.

F_0 patterns, specifically the magnitude of their rise, are also sensitive to the intonation contour on which they ride, generally rising slightly higher on marked contours, everything else being equal. If we disregard contextually induced focus or emphasis, we are left with F_0 patterns in prosodic stress groups, whose manifestation is entirely predictable from the wider context, early or late, higher or lower on the (more or less declining) intonation contour. It is hard to say to what extent this variation is voluntary and directly speaker controlled or should be ascribed to general speech production principles, which reduce articulatory explicitness through time or in unmarked versus marked contexts. But irrespective of the output control mechanism and even though F_0 patterns are an integral part of a model of speaker performance, they are not part of the abstract underlying representation, properly speaking. Stress belongs there, but its phonetic manifestation does not. Being contextually determined, however, does not deprive the predictable F_0 pattern variation of any perceptual relevance. It becomes the norm against which noneven distribution of prominence is evaluated.

Similar views about intonation as a layered structure are to be found in Bruce (1977), Fujisaki, Hirose, and Ohta (1979), Gårding (1979), ’t Hart and Collier (1979), Möbius (1993), Öhman (1968), and Vaissière (1983). Note, however, that the Danish model presented in section 2.4.10 and figure 2.10 is not simply an additive “overlay” model. The intonation contour determines the execution of the prosodic phrase contour, which in its turn sets the scale for the prosodic stress group pattern, that is, *superposition* and *subordination* are equally important aspects of the model. (The full documentation for the descriptive adequacy of the model does not find room here, but it can be found in previous work, all of it referenced and most of it summarized in Grønnum 1992.)

2.4.10 Modeling Short Utterances

Having done away with most of those parameters that are integral parts of other languages (and thus Standard Central Swedish)—sentence accents, boundary tones, and local final cues to modality, as well as final lengthening—we are left with the manifestation of stress and the overall, global intonation contour (see figure 2.10).

2.4.10.1 Utterance intonation contours The central assumption, borne out by prior acoustic analyses, is that the stress group patterns have no autonomous status other than as the manifestation of stress. They are superposed on and subordinate to the intonation contour where they occur, rises being higher early than late and higher on less falling contours, everything else being equal. The stressed syllables are frequency scaled independent of the presence or not of unstressed syllables. Furthermore, F_0 patterns are truncated rather than compressed when time is short (see figures 2.1 and 2.9), which means that there is no upward deflection of F_0 in a stress group without posttonics. And in a string of, say, four stressed syllables, as in, for instance, *Lis så Pers film* (Lis watched Per’s film), the stressed syllables are frequency scaled in the same way as in *Lissi kendte Peters tante* (Lissi knew Peter’s aunt), but their timing will be different. Carry this observation over to the model in figure 2.10. If you lift out one or more dots from a stress group, you should cut back the thin full line accordingly.

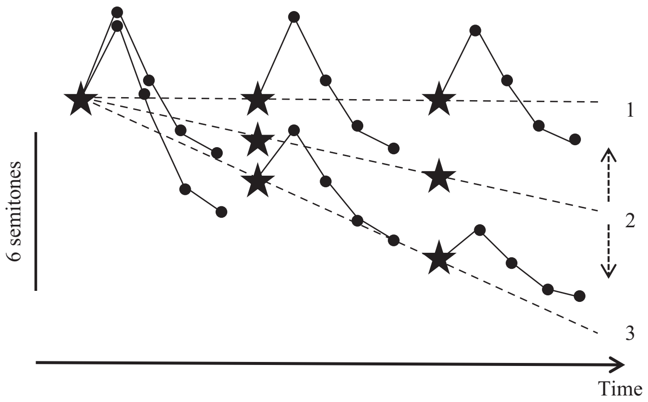


Figure 2.10
A model for the course of fundamental frequency in short utterances in Standard Copenhagen Danish. Stars represent stressed syllables and dots unstressed ones. The solid lines depict the F_0 patterns associated with prosodic stress groups, and dashed lines denote the intonation contours proper: 1, syntactically unmarked questions; 2, interrogatives with word order inversion and/or interrogative particle (variable), and nonterminal clauses; 3, terminal declaratives.

Intonation contours vary between most steeply declining—in declarative sentences used conventionally (line 3 in figure 2.10), and level—in utterances that are not marked lexically or syntactically for their interrogative function (line 1 in figure 2.10). Other questions and nonfinal clauses fill in the intermediate space, with a clear tendency for a trade-off between lexical/syntactic markers and intonation contour slope (line 2 in figure 2.10). The steepest contours typically span four to six semitones. The choice of contour slope for a given utterance is determined by syntactic and, not least, pragmatic factors, in accordance with principles of markedness and typicality, as I have suggested. Syntactically unmarked declarative sentences used conventionally are associated with unmarked intonation contours, and they are the most steeply declining of all. That makes any less falling contour marked. Typical intonation is the contour that accompanies any given sentence type when it is used conventionally. Thus, a conventional yes/no question will have a slope somewhere between unmarked (most steeply declining) and level, which means it is marked but typical. Any deviation from the typical intonation will be contrary to conventional use and have implications for illocutionary force. Thus, the most strongly marked intonation contour, line 1 in figure 2.10, will, by virtue of its nontypicality, turn a declarative sentence into an interrogative speech act.

Between the unmarked declining contour (line 3 in figure 2.10), and the level (line 1), there is room for considerable individual variation. Some speakers cluster their interrogative and nonterminal contours close to the unmarked declarative; others do the opposite and produce them just below line 1 in figure 2.10. In either case, an ordering among contours can still be observed in a trade-off with syntax: the more steeply declining marked contours typically accompany *wh*-questions, whereas the least declining contours are associated with questions with word order inversion, reminiscent of Bo's (1933) observation.

For a given intonation type, the range spanned by the contour is constant, whether it contains two or (up to) five stress groups. If it is not level, it descends gradually and evenly through stressed syllables which are equidistantly spaced in frequency, with frequency intervals that are inversely proportional to the number of stressed syllables, as depicted in the stylized contours in figure 2.11. But their timing will depend on stress group length (in terms of the number of unstressed syllables), as hinted by the nonuniformly sloping lines between neighboring stars in the leftmost and middle contours in figure 2.11.

2.4.10.2 Intonation cues to modality are global, not local It is explicit in the model in figure 2.10 that no specific and elaborate F_0 movement is attached to the end of the contour, whether to mark a default sentence accent, a prosodic boundary, or utterance modality. Intonational markedness (less or more declarative-like) and utterance completion are inherent in the global course of the intonation contour, possibly in conjunction with the derived variation in the magnitude of stress group pattern rises, as well as diminishing intensity. This property is shared by the majority of regional Danish variants.

2.4.10.3 Syntactic structure in simple sentences is immaterial Within a simple sentence, the syntactic boundaries leave no trace in the course of F_0. The onset of each stressed vowel sets off a new F_0 pattern. You can move the constituents around as shown here (translated literally), and each of the resulting utterances will have basically the same course of F_0:

Der går mange busser fra Tiflis (There go many buses from Tiflis).
Fra Tiflis går der mange busser (From Tiflis go there many buses).
Busser fra Tiflis går der mange af (Buses from Tiflis go there many of).

2.4.10.4 Longer utterances and prosodic boundaries Some speakers manage to read a sentence with eight stressed syllables on one gradually declining intonation contour, as in the bottom panel of figure 2.5. But it is more common that above four or five stressed syllables, prosodic boundaries will be inserted, creating a succession of prosodic phrase contours within the intonation contour, as in the long utterance in figure 2.12a: _R**a**sk er s**i**kkert den **e**neste d**a**nske spr**o**gforsker hvis // n**a**vn som v**i**denskabsmand er k**e**ndt // **u**den for sn**æ**vre spr**o**gvidenskabelige kr**e**dse_ (Rask is probably the only Danish linguist whose name as a scientist is known outside of narrow linguistic circles). (The double slashes mark

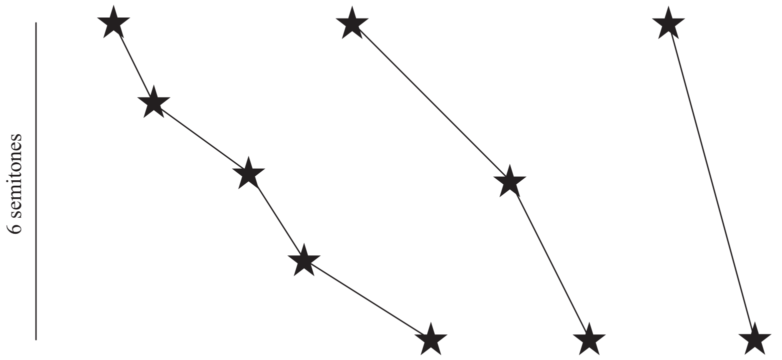


Figure 2.11
Stylized intonation contours with five (left), three (middle), and two (right) stressed syllables.

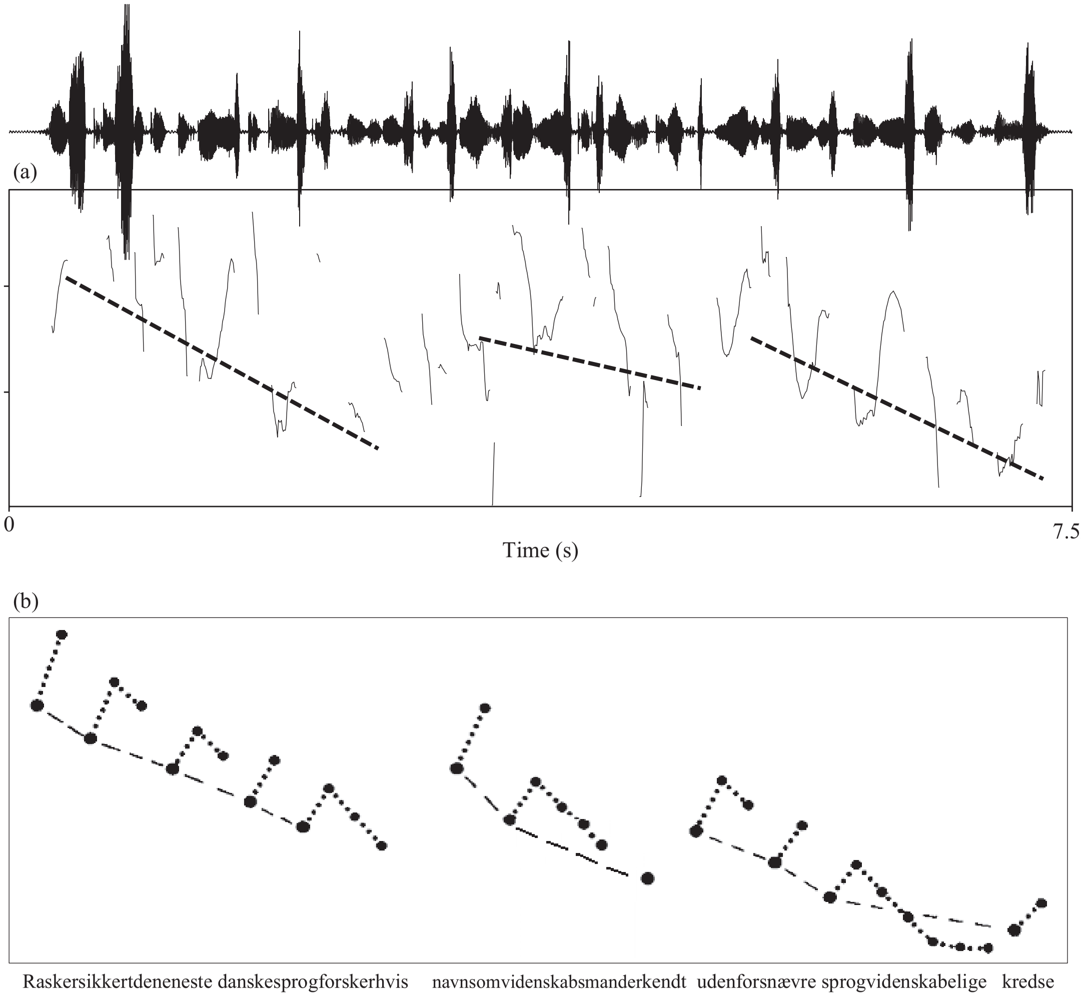


Figure 2.12

(a) Fundamental frequency tracing of a long utterance, *R**a**sk er s**i**kkert den **e**neste d**a**nske spr**o**gforsker hvis // n**a**vn som v**i**denskabsmand er k**e**ndt // **u**den for sn**æ**vre spr**o**gvidenskabelige kr**e**dse* (female speaker). (b) Stylized version of the fundamental frequency tracing in panel (a). Heavier dots denote stressed syllables and lighter dots unstressed syllables; broken lines denote the phrasal contours; and dotted lines the prosodic stress group patterns.

prosodic phrase boundaries.) Microprosody and perturbations due to the laryngealization in stød interfere in the F_0 course of the utterance, and to clarify, figure 2.12b depicts a stylized version of the utterance. This is not one smoothly declining intonation contour. Rather, the utterance has been parsed into three prosodic phrases with associated phrase contours (the broken lines), with partial resettings before the second and third phrases. Note that phrase contour onsets, defined as the frequency location of the first stressed syllable in the phrase, decline gradually through the utterance, as do offsets, defined as the frequency location of the last stressed syllable in the phrase.

One aspect of utterance-internal prosodic phrase boundaries is not in evidence in figure 2.12 (and not explicit in figure 2.13 either). The F_0 rise in the last stress group before an internal prosodic phrase boundary is slightly higher (by less than one semitone) than preceding F_0 rises in the phrase. It may be an independent gesture to signal noncompletion if it is not simply an assimilation to the higher onset of the succeeding prosodic phrase.

In the scripted material from which I derived the model of intonation presented here, the following is valid for *simple* sentences, in isolation or in combination:

1. The syntactic structure of short, simple sentences is not reflected in their intonation contour. Neither the order of constituents nor their internal structure matter.
2. Longer utterances are produced in a descending sequence of declining prosodic phrases, but the conditions governing the location of the breaks are complex: (i) A prosodic phrase must contain at least two stress groups. (ii) Prosodic phrases tend to be of equal length. But (iii) this balancing tendency is easily overruled: The prosodic boundary cannot occur within a simple first-order syntactic constituent or between semantically coherent syntactic constituents. Accordingly, *Der vil gå mange store Røde Kors busser til Grosny i Tjetjenien i aften* (Many big Red Cross buses will depart for Grozny in Chechnya this evening.) will be read (at least by fluent readers/speakers) in one sweep, in spite of its length and in spite of the formal syntactic boundary before the place complement, because buses are intimately associated semantically with their destination (or their point of departure). Thus, buses, or bus routes, may be named after their destination in one word, a compound with one main stress on the location and reduced stress on *bussen*, as in *Rombussen* (the bus to Rome).
3. Coordinate main clauses are more likely than subordinate and main clause constructions to be separated by a resetting of the intonation contour.
4. Individual intonation contours are steeper, even the text-final one, and demonstrate a greater amount of resetting between them in a succession of terminal declaratives than in a corresponding string of coordinate main clauses.

From multiple linear regression analysis of read sentences of diverse syntactic structure, consisting of up to three prosodic phrases, Petersen (1999, 2001) formulated mathematical expressions whose predictions emulated observed phrase and utterance intonation contours, confirming the global nature of utterance modality and the look-ahead involved in producing intonation contours in utterances of varying length.

2.4.11 An Extended Model

All of the observations noted thus far can be assembled into a more elaborate model, as in figure 2.13. The two Ts designate the overall textual contour (dotted line); U1, U2, and U3 are utterance contours (full lines), the first one of which is depicted with three potential slopes, 1, 2, and 3, corresponding to the contours in figure 2.10. P1 and P2

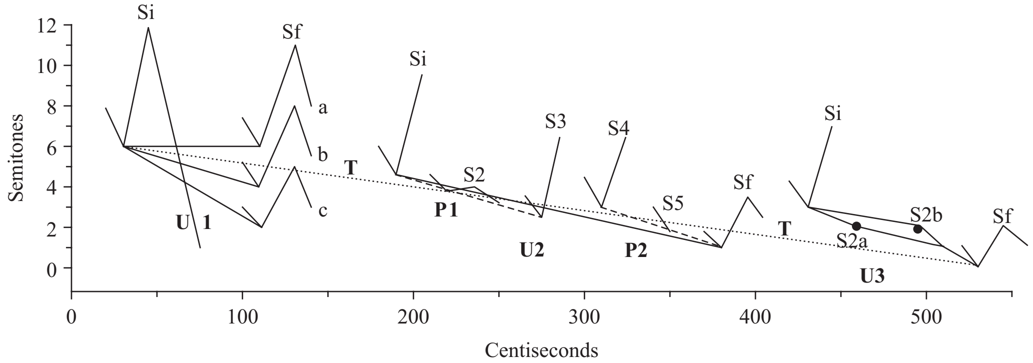


Figure 2.13
An extended model of Danish intonation.

are phrase contours (broken lines) within U2. The Ss designate stress group patterns (nonsaturated lines), where Si is initial in the utterance, Sf is final, and intermediate stress groups in the utterance are numbered successively from S2 and upward. The lack of F_0 rise in S2 (in P1/U2) is indicative of secondary stress, and S5 (in P2/U2) is a monosyllabic stress group. S2a and S2b in U3 illustrate how different duration of the initial stress group will produce different utterance contour shapes.

2.4.11.1 The textual contour Short texts are characterized by a gradual global declination, the textual contour, T, in figure 2.13. Its onset is defined by the first stressed syllable, Si, in the first utterance and its offset by the last stressed syllable, Sf, in the last utterance. It typically spans half an octave. Superposed on T are the individually declining utterance contours, U1, U2, and U3. The resetting of utterance contour onsets is only partial, so that utterance onsets lower gradually through the text and typically span three semitones. Utterance offsets also lower gradually but typically span only two semitones, so that utterance slopes are slightly steeper initially than finally in the text, everything else being equal (i.e., in utterances of equal length and identical modality). In texts of more than four utterances, the medial part of the textual contour will level out. It is unreasonable to expect speakers to be able to manipulate successive utterance onset descent to a degree finer than one semitone, particularly in view of their relatively large mutual temporal distance. Coordinate main clauses are less slanted relative to the global textual contour than are a sequence of terminal declaratives (not depicted in figure 2.13).

The organization of the phrase contours, P1 and P2, relative to the superordinate utterance contour, U2, mirrors the organization of utterances within the text. Phrases descend along the utterance contour. Each phrase is associated with its own slope, so the first phrase onset is higher than the last phrase onset, and the first phrase offset is higher than the last phrase offset. Intermediate phrases have intermediate onsets and offsets, but above four prosodic phrases, the utterance contour will level out medially so as not to frustrate speakers' control over step-down magnitude between phrase onsets.

2.4.11.2 Limitations The model's strongest limitation is its restriction to speech read aloud under laboratory conditions. Furthermore, syntactic boundaries below the sentence level are not incorporated. This is not from any a priori theoretical preclusion of their relevance, but for lack of sufficient data, and in recognition of the complexity involved. Obviously, we need more research into the way syntactic boundaries manifest themselves in intonation.

2.4.12 Subordination and Superposition: Look-Ahead and Nonlocality

Why posit such a layered, hierarchical system of simultaneous, interacting, noncategorial intonational components, of varying structural and temporal scope, where larger scope components carry and set the scale for smaller scope ones? Why is implementation of F_0 events performed on the basis of upcoming as well as preceding events and sensitive also to syntactic and semantic structuring? Why not posit a computationally simpler abstract representation, in terms of a linear sequence of categorially different, noninteracting pitch accents whose manifestation is exclusively locally determined, implemented on a left-to-right basis? This is, in point of fact, the structure of the autosegmental-metrical theory (see Ladd 1996, 2008, and Arvaniti, chapter 1, this volume), a model in which there is no intonation component separate from those sequenced local pitch accent targets, and where global trends are only the result of iterative application of local downstep rules.

First, *pitch accent*, with its connotation of phonologically distinct differences, is not an appropriate term for stress group patterns in Danish because (i) there would be only one category, and it would always align in the same fashion with the segmental material; and (ii) its phonetic manifestation is predictable. In other words, a speaker cannot—at least not in the speech style represented by the model—make a choice between various types of pitch accent. And when, in spontaneous speech, the magnitude of stress group pattern rises is manipulated, to cue varying degrees of prominence through the utterance, we are dealing with a scalar, not a binary, phenomenon. Speakers are not making a choice in those circumstances either, of a particular pitch accent, from a set of phonologically distinct ones, but are simply subordinating the manifestation of stress to the demands for signaling more or less prominence. I have seen Bornholm speakers invert their slowly falling-rising F_0 patterns across a whole utterance, with a resulting change in perceived speech style or register, and I am certain a similar mechanism, a "long component" or "setting," is operative in Standard Danish. But that does not prove the existence of another pitch accent. It shows only that the manifestation of stress interacts with parameters at other levels of description.

Finally, if pitch accent were phonological, stress would not be—a proposal whose only merit is its nonconventionality. Thus, for instance, syllables with reduced stress have all the segmental characteristics of stressed syllables (full vowel quality, vowel length, segment duration, distinct articulation, stød if the phonetic and phonological conditions are satisfied) but no autonomous pitch pattern. If pitch were the distinctive factor responsible for these properties in stressed syllables, there would be no way to account for their presence in syllables with secondary stress. So again, stress group patterns are not part of the abstract representation of intonation. This is how and why, with proper scaling, the model will cover the majority of regional variants of Standard Danish as well. The principal difference among regional varieties is in the shape and the magnitude of the stress group pattern and in its alignment with segments (see figure 2.14).

The variation in shape and range of the F_0 patterns in figure 2.14 is rather astounding. Differences among the patterns are easier to capture when the governing principle in their execution is made explicit. In all of them, except Bornholm, there is only one point in the melodic fragment that is constrained in terms of its alignment with a segment in the prosodic stress group, as indicated by the arrows in the figure. This point is

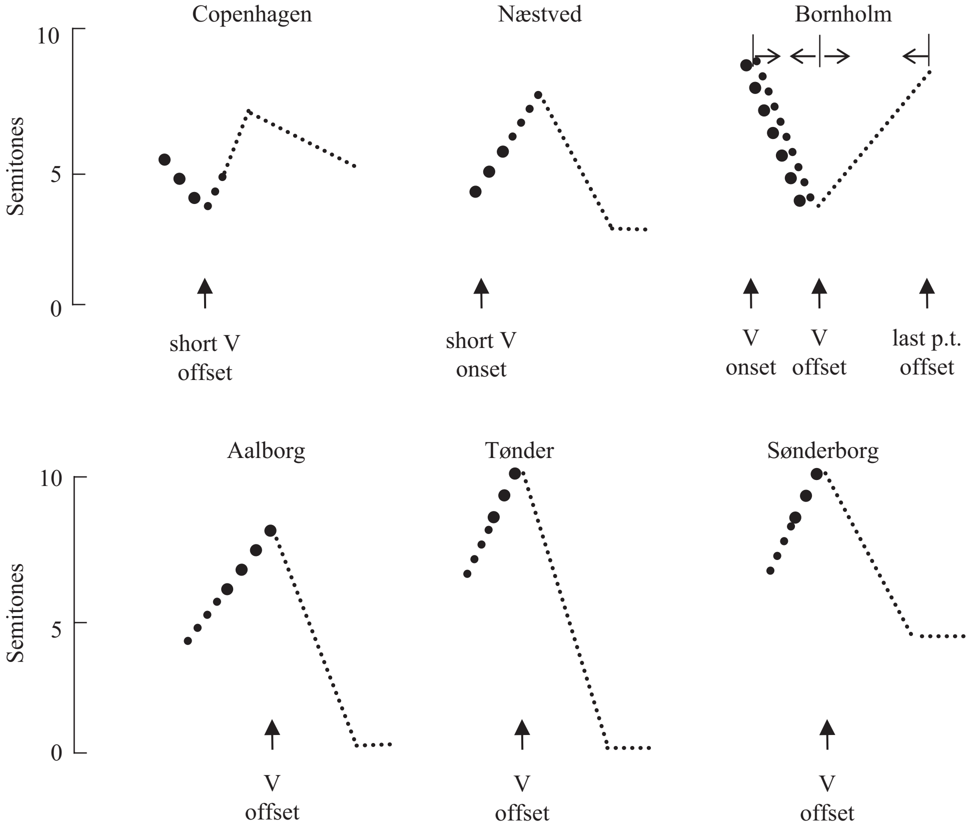


Figure 2.14
Model prosodic stress group patterns in six regional varieties of Standard Danish. Heavy dots depict the location of short stressed vowels on the F_0 pattern, medium dots depict the extension of the F_0 movement in long stressed vowels, and fine dots depict the course of unstressed syllables in the prosodic stress group. In Bornholm, the horizontal arrows enclose movements that may be expanded or compressed in time, in concordance with the duration of the stressed vowel (long or short–the falling part) and the number of posttonic (p.t.) syllables in the stress group (the rising part). Vertical arrows point to the segmental anchor point.

low relative to the first posttonic in Copenhagen and Næstved (on Zealand) and high in Aalborg, Tønder, and Sønderborg (in Jutland).

The anchor is invariably associated with the stressed vowel offset in Aalborg, Tønder, and Sønderborg, but with its onset in Næstved. In Copenhagen, the low turning point coincides with the offset of the stressed vowel if it is short and occurs about halfway through a long stressed vowel. Leaving out Bornholm, neither segments nor syllables in the prosodic stress group on either side of the anchor have separate tonal representations. In other words, they are simply strung out on the melody, like pearls of varying length on an undulating string. Voiceless segments interrupt the melody and will have microprosodic effects, but do not otherwise interfere with the score. A complete pattern materializes only insofar as there are syllables to carry it; otherwise it is truncated from the end. Accordingly, monosyllabic stress groups, with a short vowel surrounded by voiceless consonants, exhibit only that part of the melody indicated by the heavy dotted line part of these five patterns in figure 2.14.

Bornholm stands out from the other varieties by the very elastic relation between segments and F_0. The duration of the falling part of the tonal pattern varies (albeit slightly) with the duration of the stressed vowel so that the low turning point coincides with the offset of the vowel, be it short or long. The duration of the rise varies with the duration of the unstressed part of the stress group so that the rise terminates on the last posttonic syllable. This latter variation is considerable, between one and many posttonics, and perceptually the rise is more conspicuous than the fall. In other words, the pattern has three targets: a high onset, a low turning point, and a high offset. There is, however, a limit to how fast the fall and rise may be executed, a compressibility maximum; when this limit is reached—as in a monosyllabic stress group with a short vowel surrounded by voiceless consonants—the fall disappears, and what remains is the rise from a low onset to a high offset.

Whatever the F_0 patterns, they are superposed on a contour defined by the stressed syllables without interfering with their frequency scaling, but the intonation contour (its location in the text, its slope, its length), on the contrary, is decisive for the manifestation of stress group patterns. Furthermore, the manifestation of prominence degrees above normal stress demonstrates yet more syntagmatic interaction. Stress group patterns are reduced before emphasis for contrast, but not before a mere focus, that is, the realization depends on the nature of the succeeding stress group.

Furthermore, utterance and text intonation contours cannot be computed on a purely left-to-right basis either. Slope varies in inverse proportion to length. Stressed syllables in an utterance are more closely spaced in frequency in longer than in shorter utterances. In other words, the location of a stressed syllable in the frequency range is sensitive to the number of succeeding ones. For example, the second stressed syllable is higher in an utterance of five than one of three stressed syllables, everything else being equal, as depicted in figure 2.11. A similar look-ahead is operative in the combination of utterances into prosodically coherent texts. In this manner, utterance and text contours are subject to compression and expansion in time, a process that is inconceivable without look-ahead. Pre-planning also turns up in the difference between coordinate main clauses, which are less slanted relative to the global textual contour than a sequence of terminal declaratives. Look-ahead and interaction turn up in the temporal structure as well: speaking rate is somewhat accelerated before a (noninitial) focused item.

Finally, intonation contours, except the maximally marked ones, are always associated with some degree of declination. Thus, for instance, there are no conventional declaratives without declination.

These facts should make an account of sentence intonation in Danish, in terms of (varying degrees of) local downstep or range reduction, triggered by certain pitch accent configurations, perhaps a formal possibility but empty of the significance it carries in tone languages. Note that the global character and the hierarchical principle underlying the superposition and subordination do not exclude the existence of features of a singular and noninteracting character such as boundary signals. Thus, there are Danish regional varieties that definitely have a specific, final boundary tone that does not interact with what precedes it, as in Sønderborg and Bornholm. Likewise, a hierarchical organization does not preclude lexical tonal differences.

To sum up, the hierarchical and sequential models do not differ in the acknowledgment of look-ahead, but they differ in its representation. For example, in linear sequence models, longer utterance onsets are higher than shorter ones, but pitch accent relations are unaffected. In the hierarchical approach, utterance onsets need not vary, but the declining slope is less steep in longer than in shorter utterances. Liberman and Pierrehumbert (1984) stated that a pitch accent can only look back to a previous pitch accent, a phrase to a previous phrase, and so on, and apparent instances of anticipation should be explained by, for example, feature spreading or temporal overlap in the realization of the segments in question (which is really only passing the buck). The hierarchical representation entails a more direct interaction between subconstituents and superordinate structures.

2.4.13 Intonation and Meaning

The model described here has only two components: the prosodic stress group pattern and the superordinate intonation contour, with its potential subordinate phrase contours. The prosodic stress group pattern lends prominence to stressed syllables, but has no meaning in and of itself. The only potential candidate, then, for a direct link to meaning, albeit of a nonlexical nature, is the intonation contour. And indeed, figure 2.10 pointed to a correlation between intonation contour slope and sentence type. For instance, a declarative sentence in conventional use is typically associated with the most declining intonation contour: one that falls to the bottom of the speaker's normal speaking range. At the other extreme, declarative sentences that function as interrogatives are typically associated with a nondeclining intonation contour. But that is not tantamount to saying that any of these contours, in and of themselves, carry meaning in any narrow sense, at least not in linguistically coded messages. No one intonation contour invariably "means" one thing only, or invokes just one modality, just as no one modality is invariably associated with any one intonation contour. The larger context determines how any given sentence, with any given intonation contour, is to be interpreted, taking also unconventional and nontypical usage into account.

Some examples follow in table 2.1 from a corner of a phonetician's universe, all about what phoneticians do and what it takes to be one. They involve questions versus statements, and they demonstrate that any sentence, with any intonation contour, may be interpreted as either a question or a statement, according to the wider context. The numbers in parentheses in the intonation contour column refer to the contour slopes in figure 2.10.

Here are staged but plausible wider contexts for two of the examples in table 2.1. (i) The declarative sentence that opens an interview: A famous phonetician, now deceased, is invited to be interviewed on the Danish radio. The journalist opens the interview with a statement, with a typical unmarked statement intonation contour: *Du kalder dig selv fonetiker* (You call yourself a phonetician). But the utterance is correctly interpreted by the interviewee as a request for confirmation and an invitation for her to talk about herself and her life as a phonetician. (ii) The interrogative sentence that is interpreted as a polemical statement: A historian of science is being interviewed about various prominent figures in the humanities in Denmark and calls one of them a linguist, but then goes on to mention her important analysis of the acoustics of stop consonants. The journalist responds with a question: *Hvad er en fonetiker?* (What is a phonetician?) with a typical question intonation contour, but is being polemical and actually stating that with that kind of activity, one is a phonetician, not a linguist. Readers are invited to use their imagination to create contexts for the remaining six examples.

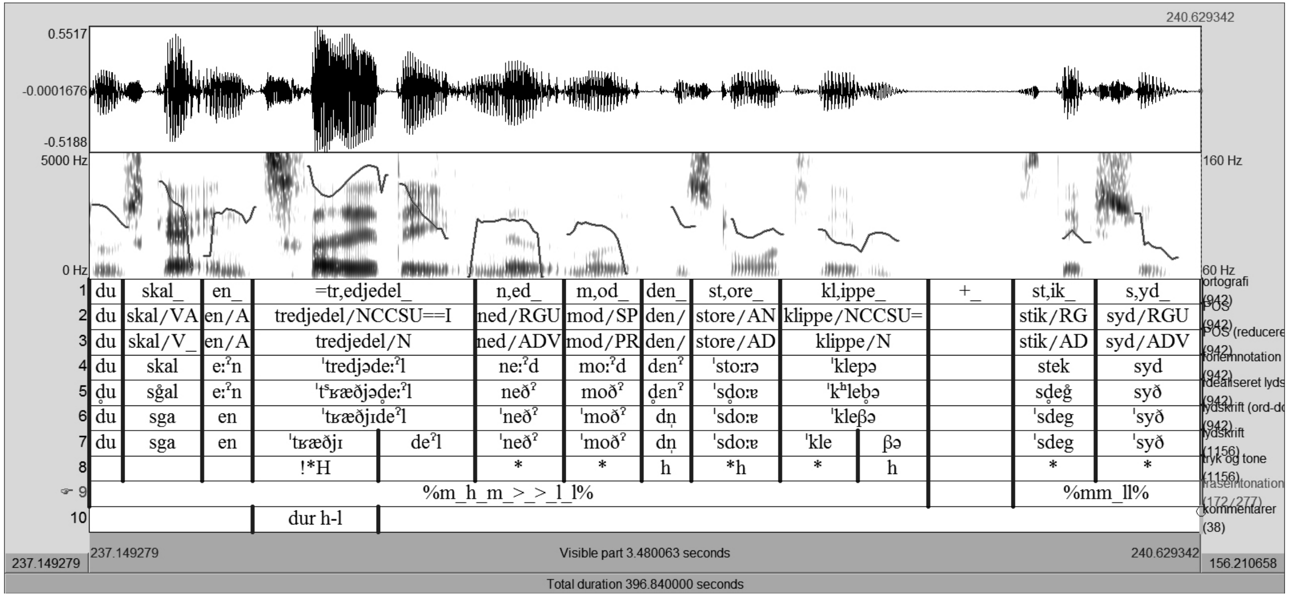
Table 2.1
Examples of declarative and interrogative sentences with different intonation contours and their interpretation in isolation and in various staged contexts

	The spoken utterance		
The sentence	Intonation contour	Interpretation in isolation	Interpretation in various contexts
Declarative: *Du kalder dig selv fonetiker* (You call yourself a phonetician.)	Nondeclining (1)	Question	*Question*: Why not a linguist? *Ironical statement*: You are NOT a phonetician.
	Declining (3)	Statement	*Question*: Opening an interview *Statement*: And yet someone else might call you a LINGUIST.
Interrogative *Hvad er en fonetiker* (What is a phonetician?)	Nondeclining (2)	Question	*Question*: True request for information. *Polemical statement*: That makes her a PHONETICIAN.
	Declining (3)	Statement	*Polemical question*: If SHE is not a phonetician, then what IS a phonetician? *Statement in conclusion*: That makes her a phonetician.

2.4.14 Transcription

The model was developed to account for speakers' production of short utterances and texts, in a distinct and formal style of speech, not with any specific application to transcription in mind, whether in fieldwork or didactic application. But its basic tenets—the stress group pattern and the intonation contour—have served well as anchor points in the transcription of spontaneous Danish speech in the DanPASS corpus (Grønnum 2009).

The corpus is segmented down to word and syllable levels wherever possible. Orthography in tier 1 in the Praat screen shot in figure 2.15 (Boersma and Weenink 2006) is supplemented with information about stress, in the shape of commas before the vowels of the stressed syllables; word class information is fully specified in tier 2 and simplified in tier 3; the phoneme representation is in tier 4; an idealized pronunciation in tier 5; the actual phonetic transcription, relatively fine grained, at the word level in tier 6; and broken into syllables in tier 7. Tiers 8 and 9 hold the prosodic information. Stressed syllables are marked with a star. If more than normally/neutrally prominent, the star is preceded by an exclamation mark. The magnitude of the rise to the first posttonic syllable may be higher than normal (H/; not evidenced in figure 2.15), simply high (H), or smaller than normal (h). If a posttonic syllable, for whatever reason, is lower than the preceding stressed syllable, it is marked in analogous fashion with L\ (much lower), L (lower), and l (a little lower), respectively (not evidenced in figure 2.15 and not very common). In tier 9, the perceived pitch level of each stressed syllable in


240.629342
0.5517
-0.0001676
-0.5188
5000 Hz
0 Hz
160 Hz
60 Hz
237.149279
237.149279
Visible part 3.480063 seconds
240.629342
156.210658
Total duration 396.840000 seconds

1	du	skal_	en_	=tr,edjedel_		n,ed_	m,od_	den_	st,ore_	kl,ippe_		+_	st,ik_	s,yd_	ortografi (942)
2	du	skal/VA	en/A	tredjedel/NCCSU==I		ned/RGU	mod/SP	den/	store/AN	klippe/NCCSU=			stik/RG	syd/RGU	POS (942)
3	du	skal/V_	en/A	tredjedel/N		ned/ADV	mod/PR	den/	store/AD	klippe/N			stik/AD	syd/ADV	POS (reducere (942)
4	du	skal	eːˀn	ˈtredjədeːˀl		neːˀd	moːˀd	dɛnˀ	ˈstoːrə	ˈklepə			stek	syd	fonemnotation (942)
5	d̥u	sg̊al	eːˀn	ˈtˢʁæðjəd̥eːˀl		neðˀ	moðˀ	d̥ɛnˀ	ˈsd̥oːɐ	ˈkʰleb̥ə			sd̥eg̊	syð	idealiseret lyds (942)
6	du	sga	en	ˈtʁæðjɪdeˀl		ˈneðˀ	ˈmoðˀ	dn̩	ˈsdoːɐ	ˈkleβə			ˈsdeg	ˈsyð	lydskrift (ord-d (942)
7	du	sga	en	ˈtʁæðjɪ	deˀl	ˈneðˀ	ˈmoðˀ	dn̩	ˈsdoːɐ	ˈkle	βə		ˈsdeg	ˈsyð	lydskrift (1156)
8				!*H		*	*	h	*h	*	h		*	*	tryk og tone (1156)
9	%m_h_m_>_>_l_l%												%mm_ll%		fraseintonation (172/277)
10				dur h-l											kommentarer (38)

Figure 2.15
Praat screen shot from the DanPASS corpus. *Source*: http://www.danpass.hum.ku.dk.

the phrase is annotated to three degrees: high (h), mid (m), and low (l), respectively. Gradual decline over several stressed syllables, for instance, from mid to low, is marked with arrows, as shown in the figure. No attempt has been made to relate contours across phrase boundaries, but the information may easily be supplied if required.

Note that any similarity with the tones and break indices (ToBI) convention for transcribing prosody (see Jun, chapter 4, this volume) is merely superficial. For the description of Danish intonation, the phonological assumptions behind ToBI are inappropriate, and as a phonetic transcription system, it is not sufficiently fine grained. (For a more detailed discussion, see Thorsen 1985, 1986a, 1986b, and Grønnum 1995.)

2.4.15 Psychological Reality

A linear representation is computationally somewhat simpler, and thus easier to implement in rule synthesis, than layered structures involving look-ahead, but that is not necessarily indicative of how speakers and listeners process intonation. And look-ahead is definitely operative and evidenced in other aspects of speech, so why not in intonation?

We are all both speakers and hearers, and although we do not a fortiori produce and perceive in the same terms, it is at least not an unlikely assumption. I hypothesize that utterance intonation in Danish is holistically conceived, and that this concept is cognitively simpler to the speaker and listener than the summation of its atomic elements, that is, the local ups and downs. Anecdotal evidence hints that this hypothesis would be worth testing. When linguistically naïve Danes are asked to characterize the melody of various Danish utterances, they typically provide gross overall shapes. They have to be pushed hard—and you have to exaggerate the patterns—for them to hear that there are local pitch movements associated with stressed syllables. The local humps are not conceived as part of the melody, and listeners seem to disregard the contribution to pitch that comes from stress.

2.5 From Read to Spontaneous Speech

What happens when the model is confronted with spontaneous speech? I do not expect look-ahead and preplanning to be much in evidence. If the speaker does not know in advance what he is going to say, but proceeds as his thoughts develop, he may not only have to insert prosodic phrase boundaries and pauses in suitable places, but he may also be forced to lower the utterance offset in longer utterances relative to shorter ones. Nor do I expect to encounter the even distribution of prominence that characterizes isolated read utterances. On the contrary, prominence will vary among the stressed words as the pragmatic context dictates. Therefore, F_0 rises in stress group patterns will vary in magnitude, with higher rises after more prominent syllables (see section 2.4.5; Jensen and Tøndering 2005a, 2005b), and global contours will not progress in smooth and regular fashion because prominence beyond neutral may propel the stressed syllable upward in the frequency range, making the contour bumpy. On the other hand, I would still expect utterance modality to be inherent in the global trend of its contour and stress group F_0 patterns to be superposed on the contour without interfering with its execution, except for the timing inherent in the duration of individual stress groups.

A handful of studies of spontaneous Danish address these issues. Dyhr (1993) analyzed a total of 302 utterances from three speakers, interviewed in their homes. He mostly confirms the validity of the model's assumptions but also finds that longer declarative utterance contours may level out toward the end or perform a rising-falling movement from start to finish. Dyhr (1995) notes that emphasis may be achieved not only by lifting the stressed syllable upward in the frequency range, but occasionally also by lowering it, relative to surrounding stressed syllables.

Grønnum and Tøndering (2007) analyzed three hundred questions of various types, plus 51 declarative utterances for comparison, from some of the dialogues in the DanPASS corpus. In spite of considerable variability in the material, there is nevertheless a tendency for different global trends among the contours, *and* an association between global trend and utterance modality, although with a slightly different ordering than in read speech. Questions with word order inversion are slightly less declining than declarative questions. Surprisingly, final F_0 rises were somewhat higher in the declarative utterances and smaller on the less declining interrogative contours, contrary to read speech. However, new information is typically found last in declaratives but first in questions (Lambrecht 1994), which may account for the distribution of F_0 rises in these dialogues.

Tøndering (2008, 2010) provides a thorough analysis of intonation and syntax in nearly all the monologues in DanPASS: 2,604 prosodic phrases, from fifteen speakers, containing between zero and fifteen stressed syllables. He encountered 550 different sequences of the four symbols in the annotation (*h*, *m*, *l*, >). For example, there are 469 prosodic phrases with four stressed words; they occur with 91 different contours (of 111 theoretically possible ones). No model can be expected to predict this kind of variability, and 'a model of intonation in spontaneous speech' is actually a contradiction in terms. But we may inquire which features of read speech recur in spontaneous speech, though we may not be able to quantify them accurately. Tøndering finds no evidence of look-ahead and preplanning in the data, in terms of equidistantly frequency scaled stressed syllables in contours of constant range. On the contrary, longer phrases have lower offsets than shorter ones do, as expected. Nor is the magnitude of the F_0 rise in stress group patterns found to decrease gradually through the phrase, also as expected. On the other hand, Tøndering reports no evidence that F_0 patterns per se—by virtue of their peak or offset frequency—have any role in the scaling of the stressed syllables. And in spite of the rather forbidding variability, he nonetheless discovers an overall declining trend in the bumpy intonation contours. (There are hardly any questions in these monologues.) He finds, furthermore, that F_0 rises are highest in the last stress group in the phrase. That is where new information is located, hence with more prominence, hence higher F_0 rises.

In summary, applied to the description of intonation in spontaneous speech, with its contextually and pragmatically induced variation in prominence, the model needs to adjust the smooth progression of the stressed syllables in prosodic phrases and utterances, as well as the gradual reduction of the associated F_0 patterns. Insofar as final stressed syllables are the more prominent ones in declaratives, that will, in and of itself, make final F_0 rises higher than preceding ones and also higher than in nondeclaratives, where new information tends to occur early, not late as in declaratives.

2.6 Conclusion

If any feature in Danish prosody can lay claim to fame for its complexity, it would be the stød (Basbøll 2005; Grønnum, Vazquez-Larruscaín, and Basbøll 2013), not intonation, which could hardly be simpler. What most obviously distinguishes standard Danish intonation from other languages—notably its nearest neighbor, Swedish—is the small inventory of components. There are only two: a global utterance contour, which may be partitioned into a succession of subordinate prosodic phrase contours and may also combine to form overarching textual contours, and a local F_0 pattern associated with the prosodic stress group. In spite of its considerable variability, intonation in spontaneous speech can be accounted for with the same two entities. The difference in final F_0 rise magnitude between declaratives and interrogatives could have been a candidate for a separate cue to modality if it were not for the contraintuitive fact that these rises are more extensive in declaratives than in interrogatives. This difference has to do not with modality per se but with the different way information is structured in questions and statements.

The other distinguishing feature is what in read speech appears as a hierarchical organization of the two components: F_0 patterns are superposed on and subordinate to the intonation contour. Superposition means that stress group patterns do not interfere with the intonation contour as defined by the stressed syllables. Subordination means that the magnitude of F_0 excursion in the stress group pattern depends partly on the kind of contour on which it rides (higher rises on less declining contours) and partly on its location on the contour (rises shrink from start to finish). The need to highlight some words in spontaneous speech and relegate others to the background introduces a gradation of acoustic and perceptual prominence that wreaks havoc with the smooth intonation contours of read speech. Furthermore, with more or less prominence, F_0 patterns are more or less extensive. Accordingly, they do not shrink gradually toward the end, and their subordination to the utterance contour is not discernible. But as long as the frequency scaling of the stressed syllables is unaffected by the associated F_0 patterns in terms of their peak or offset frequency, the intonation contour is still superordinate to the manifestation of the prosodic stress group pattern, and a hierarchy can be said to exist.

Evidently, the last word has not been said about Danish intonation, and many more analyses of spontaneous speech are awaiting future investigations.

Acknowledgments

Many thanks to Hans Basbøll for extensive and insightful comments on a draft of this chapter.

Note

1. Until the mid-1980s I published under my married name as Nina Thorsen and until 1989 as Nina Grønnum Thorsen.

References

Andersen, P. 1954. "Dansk Fonetik." In *Nordisk Lærebog for Talepædagoger*, 308–354. Copenhagen: Rosenkilde og Baggers Forlag.

Arnholtz, A. 1939–1940. "Talekunstens og talelærens udvikling og stilling i Danmark." *Nordisk Tidsskrift for Tale og Stemme* 4:82–98.

Arnholtz, A. 1942. "Om vulgær-københavnskens melodik." *Nordisk Tidsskrift for Tale og Stemme* 6:23–24.

Arnholtz, A. 1954. "Dansk Foredragslære." In *Nordisk Lærebog for Talepædagoger*, 422–451. Copenhagen: Rosenkilde og Baggers Forlag.

Basbøll, H. 2005. *The Phonology of Danish*. Oxford: Oxford University Press.

Bloch, S. N. J. 1805. *Forsøg til en Theorie for det udvortes Foredrag*. Odense, Denmark: S. Hempel.

Bo, A. 1933. *Tonegangen i Dansk Rigsmål*. Copenhagen: Branner.

Boersma, P., and D. Weenink. 2006. *Praat: Doing Phonetics by Computer*. Computer program praat5404 (accessed December 29, 2014, at http://www.praat.org/).

Brink, L., and J. Lund. 1975. *Dansk Rigsmål 1–2, Lydudviklingen siden 1840 med særligt henblik på sociolekterne i København*. Copenhagen: Gyldendal.

Bruce, G. 1977. *Swedish Word Accents in Sentence Perspective*. Lund, Sweden: Gleerup.

Dyhr, N. J. 1993. "An Acoustical Investigation of the Fundamental Frequency in Danish Spontaneous Speech." In *Nordic Prosody VI*, edited by B. Granström and L. Nord, 23–32. Stockholm: Almqvist and Wiksell.

Dyhr, N. J. 1995. "The Fundamental Frequency in Danish Spontaneous Speech with Special Reference to Syllables Boosted for Emphasis." In *Aspects of Danish Prosody*, edited by J. Rischel and H. Basbøll, 49–67. Odense, Denmark: Odense University Press.

Fischer-Jørgensen, E. 1946, 1947. Unpublished lecture notes (now belonging to the author).

Fischer-Jørgensen, E. 1983. "The Acoustic Manifestation of Stress in Danish with Particular Reference to the Reduction of Stress in Compounds." In *Sound Structures: Studies for Antonie Cohen*, edited by M. van den Broecke, V. van Heuven, and W. Zonneveld, 81–104. Dordrecht, the Netherlands: Foris Publications. A more comprehensive version appears in *Annual Report of the Institute of Phonetics, University of Copenhagen* 18 (1984): 45–161.

Forchhammer, G. 1903. *Om Nødvendigheden af Sikre Meddelelsesmidler i Døvstummeundervisningen*. Copenhagen: J. Frimodts Forlag.

Fujisaki, H., K. Hirose, and K. Ohta. 1979. "Acoustic Features of the Fundamental Frequency Contours of Declarative Sentences in Japanese." *Annual Bulletin—Research Institute of Logopedics and Phoniatrics* 13:163–173.

Gårding, E. 1979. "Sentence Intonation in Swedish." *Phonetica* 36:207–215.

Grønnum, N. 1990. "Prosodic Parameters in a Variety of Regional Danish Standard Languages, with a View towards Swedish and German." *Phonetica* 47:182–214.

Grønnum, N. 1991. "Terminality and Completion in Danish, Swedish and German." In *Proceedings of the Twelfth International Congress of Phonetic Sciences*, edited by the Congress Committee and the Organizing Committee, vol. 4, 270–273. Aix-en-Provence, France: Université de Provence, Service des Publications.

Grønnum, N. 1992. *The Groundworks of Danish Intonation: An Introduction*. Copenhagen: Museum Tusculanum Press.

Grønnum, N. 1993. "Perceptual Invariance in Danish Stress Group Patterns." In *Nordic Prosody VI*, edited by B. Granström and L. Nord, 77–85. Stockholm: Almqvist and Wiksell.

Grønnum, N. 1994. "Rhythm, Duration and Pitch in Regional Variants of Standard Danish." *Acta Linquistica Hafniensia* 27:190–218.

Grønnum, N. 1995. "Superposition and Subordination in Intonation: A Non-Linear Approach." In *Proceedings of the Thirteenth International Congress of Phonetic Sciences*, edited by K. Elenius and P. Branderud, vol. 2, 124–131. Stockholm: Stockholm Congress Organisers at Kungliga Tekniska Högskolan and Stockholm University.

Grønnum, N. 1998a. “A Critical Remark to D. R. Ladd’s Intonational Phonology.” *Journal of Phonetics* 26:109–112.

Grønnum, N. 1998b. “Intonation in Danish.” In *Intonation Systems: A Survey of Twenty Languages*, edited by D. Hirst and A. di Cristo, 131–151. Cambridge: Cambridge University Press.

Grønnum, N. 2000. “Danish: Illustrations of the IPA.” *Journal of the International Phonetic Association* 28:99–105.

Grønnum, N. 2007. *Rødgrød med Fløde. En lille bog om dansk fonetik*. Copenhagen: Akademisk Forlag.

Grønnum, N. 2009. “A Danish Phonetically Annotated Spontaneous Speech Corpus (DanPASS).” *Speech Communication* 51:594–603. https://doi:10.1016/j.specom.2008.11.002. The corpus is available at www.danpass.hum.ku.dk.

Grønnum, N., and J. Tøndering. 2007. “Question Intonation in Non-Scripted Danish Dialogues.” In *Proceedings of the Sixteenth International Congress of Phonetic Sciences*, edited by J. Trouvain and W. J. Barry, 1229–1232. Saarbrücken, Germany: Saarland University.

Grønnum, N., M. Vazquez-Larruscaín, and H. Basbøll. 2013. “Danish Stød: Laryngealization or Tone.” *Phonetica* 70:66–92.

Heger, S. 1981. *Sprog og Lyd: Elementær Dansk Fonetik*. Copenhagen: Akademisk Forlag.

Hjelmslev, L. 1951. “Grundtræk af det danske udtrykssystem med særligt henblik på stødet.” *Selskab for Nordisk Filologi. Årsberetning for 1947-49-50*:12–24. English translation by Francis Whitfield, “Outline of the Danish Expression System with Particular Reference to the Stød.” In L. Hjelmslev, *Essais linguistiques II* (= *Travaux du Cercle Linguistique de Copenhague*, vol. 14), edited by N. Ege, E. Fischer-Jørgensen, K. Togeby, and F. J. Whitfield, 1979, 247–266. Copenhagen: Nordisk Sprog og Kulturforlag.

Hjelmslev, L. 1954. “Almindelig Fonetik.” In *Nordisk Lærebog for Talepædagoger*, 233–307. Copenhagen: Rosenkilde og Baggers Forlag.

Høysgaard, J. P. 1747. *Accentuered og Raisonnered Grammatica*. In *Danske Grammatikere, Bind IV*, edited by H. Bertelsen, 1920, 249–487. Copenhagen: Nordisk Forlag. Reprint 1979, Copenhagen: Det Danske Sprog og Litteratur Selskab.

Høysgaard, J. P. 1769. *Første Anhang til den Accentuerede Grammatika*. In *Danske Grammatikere, Bind V*, edited by H. Bertelsen, 1923, 507–550. Copenhagen: Nordisk Forlag. Reprint 1979, Copenhagen: Det Danske Sprog og Litteratur Selskab,

Jensen, C., and J. Tøndering. 2005a. “Choosing a Scale for Measuring Perceived Prominence.” In *Proceedings of 6th Interspeech 2005 and 9th European Conference on Speech Communication and Technology (EUROSPEECH)*, 2385–2388. International Speech Communication Association (ISCA).

Jensen, C., and J. Tøndering. 2005b. “Perceived Prominence and Scale Types.” In *Proceedings of the XVIIIth Swedish Phonetics Conference*, edited by A. Eriksson and J. Lindh, 111–114. Göteborg, Sweden: Göteborg University.

Jensen, J. 1833. *Forsøg til en Dansk Sproglære*. Copenhagen: Den Gyldendalske Boghandel.

Jerndorff, P. 1897. *Om Oplæsning*. Copenhagen: Nordisk Forlag.

Jerndorff, P. 1911. *Tale og Læsning*. Copenhagen: Nordisk Forlag.

Jespersen, O. 1897–1899. *Fonetik*. Copenhagen: Det Schubotheske Forlag.

Jespersen, O. 1934. *Modersmålets Fonetik*. Copenhagen: Gyldendal.

Ladd, D. R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. R. 2008. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Lambrecht, K. 1994. *Information Structure and Sentence Form: Topic, Focus, and the Mental Representation of Discourse Referents*. Cambridge: Cambridge University Press.

Liberman, M. Y., and J. Pierrehumbert. 1984. "Intonational Invariance under Changes in Pitch Range and Length." In *Language Sound Structure*, edited by M. Aronoff and R. T. Oehrle, 157–233. Cambridge, MA: MIT Press.

Lützen, Peter H. 2010. *Det Sproglige i Dansk*. Copenhagen: Dansklærerforeningens Forlag.

Mikkelsen, K. 1894. *Dansk Sproglære*. Copenhagen: Lehmann and Stages Forlag.

Möbius, B. 1993. *Ein quantitatives Modell der deutschen Intonation*. Tübingen, Germany: Max Niemeyer Verlag.

Öhman, S. E. G. 1968. "A Model of Word and Sentence Intonation." *Speech Transmission Laboratory: Quarterly Progress and Status Report*: 6–11. Stockholm: Royal Institute of Technology.

Oxenvad, E. 1937. *Vort Sprog, Bind I*. Copenhagen: Nordisk Forlag.

Petersen, N. R. 1999. "Modelling Danish Sentence and Phrase Intonation." In *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, edited by J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey, 925–928. Berkeley: University of California Press.

Petersen, N. R. 2001. "Modelling Fundamental Frequency in First Posttonic Syllables in Danish Sentences." In *Proceedings of Eurospeech 2001 Scandinavia*, edited by P. Dalsgaard, B. Lindberg, and H. Benner, 939–943. Aalborg, Denmark: Center for Personkommunikation, Aalborg University.

Rischel, J. 1980. "Phrasenakzent als Signal des Objekts ohne 'determiner' im Dänischen." *Kopenhagener Beiträge zur germanistischen Linguistik, Sonderband 1*: 262–279.

Rischel, J. 1983. "On Unit Accentuation in Danish—and the Distinction between Deep and Surface Phonology." Special issue on prosody, edited by H. Basbøll. *Folia Linguistica* 17:51–197.

Sneedorff-Birch, F. 1835 *Dansk Retskrivningslære*. Copenhagen: J. H. Schubothe.

't Hart, J., and R. Collier 1979. "On the Interaction of Accentuation and Intonation in Dutch." In *Proceedings of the Ninth International Congress of Phonetic Sciences*, edited by N. Grønnum, E. Fischer-Jørgensen, and J. Rischel, vol. 2, 395–402.

Thorsen, N. 1978. "An Acoustical Investigation of Danish Intonation." *Journal of Phonetics* 6:151–176.

Thorsen, N. 1979. "Interpreting Raw Fundamental Frequency Tracings of Danish." *Phonetica* 36:57–78.

Thorsen, N. 1980a. "Intonation Contours and Stress Group Patterns in Declarative Sentences of Varying Length in ASC Danish." *Annual Report of the Institute of Phonetics, University of Copenhagen* 14:1–29.

Thorsen, N. 1980b. “Neutral Stress, Emphatic Stress, and Sentence Intonation in Advanced Standard Copenhagen Danish.” *Annual Report of the Institute of Phonetics, University of Copenhagen* 14:121–205.

Thorsen, N. 1980c. “A Study of the Perception of Sentence Intonation: Evidence from Danish.” *Journal of the Acoustical Society of America* 67:1014–1030.

Thorsen, N. 1980d. “Word Boundaries and F_0 Patterns in Advanced Standard Copenhagen Danish.” *Phonetica* 37:121–133.

Thorsen, N. 1981. “Intonation Contours and Stress Group Patterns in Declarative Sentences of Varying Length in ASC Danish: Supplementary Data.” *Annual Report of the Institute of Phonetics, University of Copenhagen* 15:13–47.

Thorsen, N. 1982a. “On the Variability in F_0 Patterning and the Function of F_0 Timing in Languages Where Pitch Cues Stress.” *Phonetica* 39:302–316.

Thorsen, N. 1982b. “Selected Problems in the Tonal Manifestation of Words Containing Assimilated or Elided Schwa.” *Annual Report of the Institute of Phonetics, University of Copenhagen* 16:37–100.

Thorsen, N. 1983a. “Standard Danish Sentence Intonation: Phonetic Data and Their Representation.” Special issue on prosody, edited by H. Basbøll. *Folia Linguistica* 17:187–220.

Thorsen, N. 1983b. “Two Issues in the Prosody of Standard Danish.” In *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd, 27–38. Berlin: Springer-Verlag.

Thorsen, N. 1984a. “F_0 Timing in Danish Word Perception.” *Phonetica* 41:17–30.

Thorsen, N. 1984b. “Variability and Invariance in Danish Stress Group Patterns.” *Phonetica* 41:88–102.

Thorsen, N. G. 1985. “Intonation and Text in Standard Danish.” *Journal of the Acoustical Society of America* 77:1205–1216.

Thorsen, N. G. 1986a. “Intonation and Text in Standard Danish: With Special Reference to the Abstract Representation of Intonation.” In *Proceedings of the Fifth International Phonology Meeting*, edited by S. L. Fulmer, W. U. Dressler, H. C. Luschützky, O. E. Pfeiffer, and J. R. Rennison, 301–309. Cambridge: Cambridge University Press.

Thorsen, N. G. 1986b. “Sentence Intonation in Textual Context: Supplementary Data.” *Journal of the Acoustical Society of America* 80:1040–1047.

Thorsen, N. G. 1987. “Suprasegmental Transcription.” In *Probleme der phonetischen Transkription*, edited by A. Almeida and A. Braun, 79–109. Stuttgart, Germany: Franz Steiner Verlag.

Thorsen, N. G. 1988. “Default Sentence Accents and Focal Sentence Accents.” In *Papers from the Second Swedish Phonetics Conference, May 1988*, edited by S. A. J. Wood, 120–123. Lund, Sweden: Lund University, Department of Linguistics.

Tøndering, J. 2008. *Skitser af prosodi i spontant dansk*. Copenhagen: Det Humanistiske Fakultet, Copenhagen University.

Tøndering, J. 2010. “Prosodiske fraser og syntaktisk struktur i spontan tale.” *Nydanske Studier* 39:166–198.

Vaissière, J. 1983. “Language-Independent Prosodic Features.” In *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd, 53–66. Berlin: Springer-Verlag.

3

A Multilevel, Multilingual Approach to the Annotation and Representation of Speech Prosody

Daniel Hirst

3.1 Introduction: Prosodic Annotation

Instinctively, when we listen to an utterance, we make a difference between what is being said, and how it is said, that is, the prosody of the utterance. There is, generally, a fairly large consensus among linguists (and nonlinguists) on the annotation of what is said. Interestingly, though, there is a remarkable lack of consensus on the annotation and representation of how it is said. This difference can partly be explained by the fact that most writing systems provide a fairly reliable annotation of what is said, whereas there is very little written indication of how an utterance is spoken, apart from a limited number of abstract prosodic characteristics that are indicated by punctuation.

The result is that there has been a bewildering variety of annotation systems for speech prosody. Until quite recently, practically every researcher has developed their own annotation system: it was quite rare for anyone to use someone else's system. Even for the British School, which had a fairly consistent approach to intonation patterns, there have been quite considerable differences among the annotation used by different authors such as Palmer (1924), Armstrong and Ward (1926), Jassem (1952), Kingdon (1958), O'Connor and Arnold (1961), Faure (1962), Halliday (1967), Crystal (1969), Couper-Kuhlen (1986), or most recently Wells (2006).

There have been at least two partial exceptions to this general tendency. One exception was the tones and break indices (ToBI) annotation system (Silverman et al. 1992), developed as a standard annotation system for American English and later applied to several different languages (Jun 2005). Another exception was my International Transcription System for Intonation (INTSINT), described in more detail in section 3.6.1. The original version of the system (Hirst 1987) was based on an inventory of minimal pitch contrasts found in published descriptions of intonation patterns of several different languages. The aim was to provide a tool for the systematic description of these intonation patterns, something on the lines of the International Phonetic Alphabet. The INTSINT system was later used for nine of the twenty-two intonation systems described in Hirst and Di Cristo (1998).

3.2 Levels of Representation

I have suggested (Hirst, Di Cristo, and Espesser 2000) that one of the reasons for this lack of consensus is that different annotation systems are in fact annotating different levels of representation. A similar point is made by Taylor (2009) who notes that different authors have completely different approaches, phonological, phonetic, acoustic:

> These different approaches are of course quite common in all areas of language study; what is of note is that intonation seems to be one of the few areas where engineering and scientific approaches still have enough commonality to constitute a single field. (244)

The prosody of an utterance can be described at several different levels. In particular, I have argued that it is important to make a distinction between prosodic forms and prosodic functions. *Prosodic forms* are how the prosody affects what the utterance sounds like. *Prosodic functions* are what the prosody does, and how it contributes to the interpretation of an utterance. I mentioned that the only written correspondence to speech prosody is punctuation, which encodes a limited number of abstract prosodic characteristics. The difference between a full stop, a comma, and ellipsis, a question mark, and an exclamation point, for example, corresponds to different ways in which the utterance is intended to be interpreted, for example:

(1) a. OK. b. OK, c. OK . . . d. OK? e. OK!

Here, the first is intended as a statement, the second as an unfinished statement, the third as a tentative statement, the fourth as a question, and the fifth as an emphatic statement. These are clearly examples of prosodic functions, and the forms by which these functions are encoded may vary considerably from language to language and even from dialect to dialect. The example of *OK* is particularly convenient for this, because the expression is used in numerous languages throughout the world, so the representations in (1) are not necessarily specific to a particular language.

This distinction between prosodic forms and prosodic functions[1] is, in my view, one of the sources of the disparity among prosodic annotation systems. In a ToBI annotation, such as, for example:

(2) a. OK
b. % H*L-L% //

the symbols H and L are clearly annotating prosodic form, while the symbols *, -, and % seem much closer to indicating prosodic functions: whether a syllable is prominent, whether it is separated from what follows. Interestingly, Wightman (2002) reported that there was considerably higher intertranscriber agreement over the place of accents and boundaries than over the nature (H or L) of those accents and boundaries. In terms of our distinction between prosodic forms and prosodic functions, this can be interpreted as showing that listeners are much more proficient at identifying prosodic functions than prosodic forms. This seems a reasonable result, because it is well known that speakers and listeners are far more proficient at performing linguistic tasks (i.e., interpreting utterances) than metalinguistic tasks, such as describing the forms of utterances.

It is also an interesting result in the light of the fact that for computers, the results are exactly the opposite: machines are excellent at describing the forms of utterances but rather poor at interpreting them.

Prosodic forms, however, do not constitute a single level of representation. In Hirst, Di Cristo, and Espesser (2000), I suggested that we need to distinguish at least three levels of prosodic form. Most abstract would be a level of underlying phonological representation, which would be the level at which the prosodic form of an utterance is linked to its prosodic function. Between this underlying phonological representation and the acoustic signal, I believe that we can usefully distinguish two more levels: a phonetic representation and a surface phonological representation.

The term *phonetic representation* has been used to cover a number of different phenomena that really need to be distinguished. Phonetics is sometimes used as a synonym for the acoustics and physiology of speech. It should be clear that I wish to distinguish these levels of analysis. *Phonetics* is also sometimes used as a synonym for *surface phonology*, as in the terms *phonetic transcription* or the *International Phonetic Alphabet*. My use of the term *phonetic* follows that of Trubetzkoy (1949), for whom the distinction between phonology and phonetics is one between discrete and continuous phenomena. In this sense, then, a "phonetic transcription" would more appropriately be termed a "surface phonological transcription."

The level of surface phonology is a level of distinctive discrete categories with which we can describe surface phenomena cross-linguistically. The level of phonetics is the level of continuously variable phenomena from which we have factored out universal constraints on the production and perception of sounds. There has been much recent discussion about the interface between phonetics and phonology. Therefore, it would seem rather that phonetics is itself an interface: the interface between cognitive representations (phonology) and the physical manifestations of sound, perception, and articulation.

Neither of these two levels of representation, phonetics and surface phonology, should necessarily be thought of as cognitive representations. It is quite possible that in a complete theory of prosodic representation, neither of these levels would be necessary. Nevertheless, both levels appear to be useful as heuristic descriptive tools, particularly in the task of describing the prosody of languages that have not yet been adequately studied.

In the rest of this chapter, I first give a brief sketch of a functional prosodic annotation system that I developed many years ago for English intonation (Hirst 1977). I then discuss the relationship of this annotation system to the notion of prosodic structure. After this, I work back up from the acoustic signal via the levels of phonetic representation and surface phonological representation, before giving some consideration to the nature of underlying phonological representation of prosody.

Most of the examples I give are in English and a few in French. These are two languages that are geographically close but prosodically rather different. There is nothing, however, in the framework I describe that is specific to either of these two languages, although it remains an empirical question whether the framework could be applied to other languages without substantial modification.

3.3 Prosodic Functions

In Hirst (1977), a revised and translated version of my unpublished PhD thesis (Hirst 1974), I proposed a system of functional annotation to account for some of the ways in which intonation contributes to the meaning of an utterance. The book, written just a few years after Chomsky and Halle's *Sound Pattern of English* (1968), proposed a functional description of English intonation by means of a set of five distinctive features for which I coined the term *intonative features*: [±STRESS; ±CENTRE; ±EMPHASIS; ±BOUNDARY; ±TERMINAL]. These features were transcribed using the following symbols:

(3) a. ‘no b. °no c. $\underline{\text{no}}$ d. /no, no / e. no+ f. no //

The symbol in (3d) was used either at the beginning of an utterance or where the terminal/ nonterminal nature of the boundary was not relevant, while (3e) and (3f) correspond, respectively, to [+BOUNDARY; –TERMINAL] and [+BOUNDARY; +TERMINAL].

These features were each justified by examining minimal pairs where only a difference of the value of the feature in question corresponded to a clear difference in meaning. In most cases, the ambiguity is not complete, and contexts can be imagined where the interpretations given here could be reversed, but it stills seems that, at least for speakers of British English, the most obvious interpretation of each utterance corresponds to the one given here and that this is triggered by the feature described.

The feature [±STRESS],[2] for example, was justified by the minimal pair:

(4) a. he ‘ate a little ‘pudding (= he ate a small quantity of pudding)
b. he ‘ate a ‘little ‘pudding (= he ate a small pudding)

The feature [±CENTRE] (also known as *intonation centre*, *nuclear stress*, *primary sentence accent*, *tonic syllable*, etc.) can be justified by the minimal pair:

(5) a. it’s a ‘good ‘job °too. (= It’s a good thing too)
b. it’s a ‘good °job, °too (= The job is also a good one)

The feature [±EMPHASIS] can be justified by the pair:

(6) a. if you °give it to me, I’ll °mend it (= I’ll mend it for you)
b. if you °give it to me, I’ll °mend it (= I’ll only mend it if you let me keep it)

The feature [±BOUNDARY] (i.e., intonation unit boundary) can be justified by the following pair:

(7) a. she ‘should have °phoned / her ‘mother was °worried (= her mother)
b. she ‘should have °phoned her / ‘mother was °worried (= our mother)

Finally the feature [±TERMINAL] can be justified by the pair:

(8) a. / ‘would you like °tea+ or °coffee+ (= or something else?)
b. / ‘would you like °tea+ or °coffee// (= which?)

My conclusion, in the presentation of these “features,” and which I still believe today, was that:

> the five intonative features we have described are sufficient to account for all the syntactic ambiguities we have so far come across and which are disambiguated by intonation, including a considerable number of ambiguities which up to now have always been treated from a semantic, “attitudinal” point of view. (Hirst 1977, 44)

This does not mean that these five features exhaust the number of different intonational meanings that can be conveyed by prosody. They do, however, to my mind, constitute a minimal “core” system of annotation for prosodic functions, which any theory of intonational meaning should be able to account for.

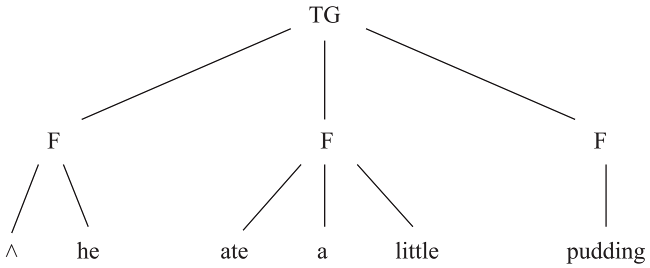
3.4 Prosodic Structure

Since I wrote my thesis and book, a lot of research has been done, particularly in the framework of autosegmental-metrical phonology, on the nature of prosodic structure (Liberman and Prince 1977, Selkirk 1984, Beckman and Pierrehumbert 1986, Nespor and Vogel 1986, Goldsmith 1990, Ladd 1996). It has today become a common point of view that both stress and boundaries are nothing but the reflection of a prosodic structure that is, at least to some extent, independent from syntactic structure. This seems a more convincing approach than a distinctive feature analysis.

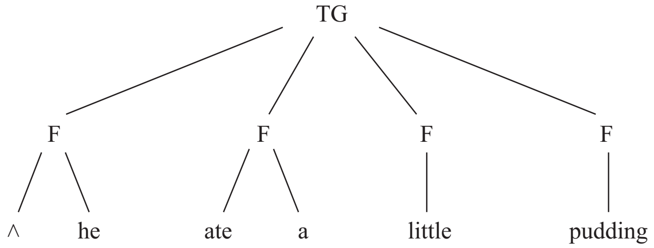
3.4.1 The Stress Foot

Halliday (1967), following Abercrombie (1964), had already suggested that an English utterance is a hierarchical phonological structure, containing four layers: the *tone group*, the *foot*, the *syllable*, and the *phoneme*. With this model, examples like (4) could be represented phonologically (omitting the level of the syllable and phoneme) as:

(9) a.



b.

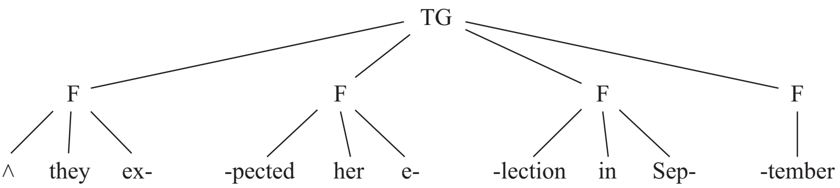


The symbol [^] in the first foot is what Halliday calls a "silent ictus," corresponding to a rhythmic beat occurring before the first syllable of the foot, just as a rest is used in musical notation for a similar purpose.

The Abercrombie/Halliday foot, often called the *stress foot* to distinguish it from the more restricted phonological unit of autosegmental-metric phonology, can be defined for speech as a sequence of syllables beginning with an accented syllable or with a silent beat at the beginning of a sentence, and continuing up to (but not including) the next accented syllable or the end of the *tone group*.

In (9), the foot boundaries all correspond to word boundaries, but this is not always necessarily the case. In (10), none of the foot boundaries corresponds to a word boundary, except the last.

(10)

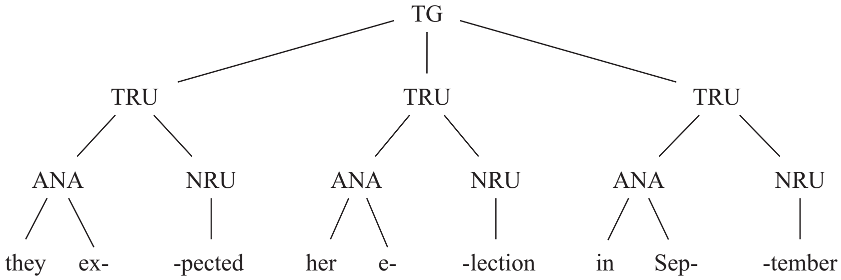


3.4.2 Tonal Units and Rhythm Units

Unlike Abercrombie and Halliday, who use the same unit to describe both melody and rhythm, Jassem (1952), more than ten years earlier, had made a clear distinction between what he called the *tonal unit*, which is conceived of as the domain of occurrence of local pitch movements in English, on the one hand, and the *narrow rhythm unit* and the *anacrusis*, on the other hand, conceived of as the domain of segmental timing and rhythm. The tonal unit, in Jassem's model, was in fact identical to the unit which was, thirteen years later, called the *foot* by Abercrombie, as defined in section 3.4.1.

The narrow rhythm unit (NRU) is similar to the foot, except for the fact that it does not usually cross word boundaries, except in cases of enclitics, which are treated prosodically as belonging to the previous word. Any syllables that are not part of a NRU form an anacrusis (ANA), in which, according to Jassem, the syllables are "pronounced extremely rapidly" (40). The ANA, together with the following NRU, together make up what Jassem termed the *total rhythm unit* (TRU). Example (10) in this analysis would look like (11):

(11)



Jassem gave several examples where his representation allowed subtle distinctions of rhythm to be made that could not be captured without the notion of ANA. In the following minimal pair, for example:

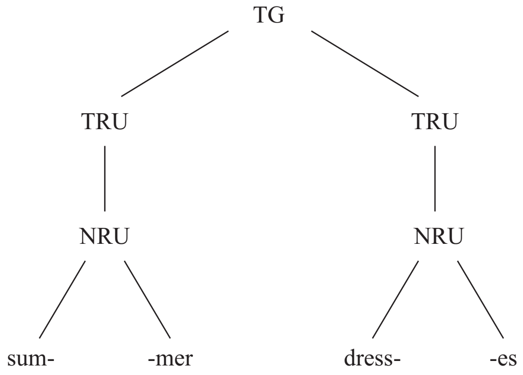
(12) a. Summer dresses b. Some addresses

Jassem notes that although the phonemes and stresses are identical, there is a subtle difference of rhythm between the two: the first syllable of (12a) is shorter than that of (12b), whereas the second syllable is longer in (12a) than in (12b). He attributes this difference to the fact that the first two syllables of (12a) constitute a single NRU, whereas in (12b) the first syllable constitutes an NRU on its own and the second syllable constitutes an ANA. He proposed to represent this in the phonetic transcription by the simple device of a space after each NRU, as in:

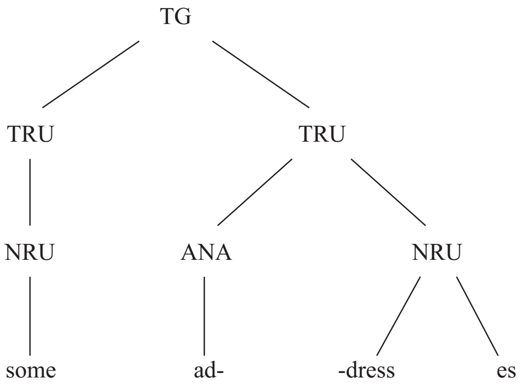
(13) a. /ˈsʌmə ˈdresiz/ b. /ˈsʌm əˈdresiz/

corresponding to the tree representations:

(14) a.



b.



In a recent collection of articles published to celebrate Wiktor Jassem's ninetieth birthday, I suggested (Hirst 2012) that Jassem's TRU does not actually play any phonological role. Instead we can combine the tonal units and the rhythm units (ANAs and NRUs) into a single representation. In fact, if we do that, we can note that there is no longer any need to make a formal distinction between ANA and NRU, because the NRU will always coincide with the beginning of a tonal unit: both can simply be characterized as rhythm units. It is here that the distinction between stress and accent plays a crucial and, I believe, particularly interesting role.

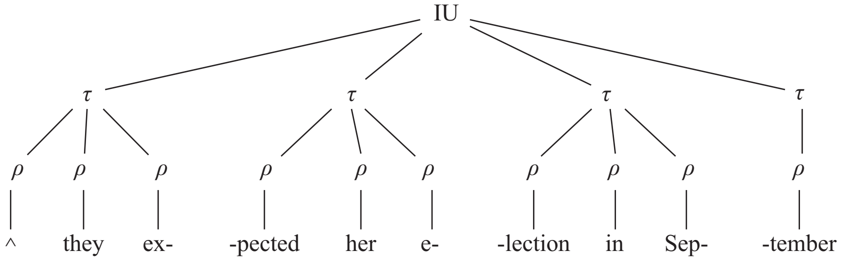
Following Bolinger (1958), I interpret *accent* as referring to the physical manifestation of prominence in an utterance and *stress* as referring to the syllable or syllables of a word that are lexically marked and normally carry this prominence when the word is accented.

In my interpretation of Jassem's model, then, the *tonal unit*, which I label [τ], begins with an accented syllable, or at the beginning of the *intonation unit*, which I label [IU] (=Jassem's tone group) and continues until the next accented syllable or until the end of the intonation unit.

Deviating slightly from Jassem's original formulation, but not, I hope, from the spirit of his analysis, I now define a *rhythm unit* (which I label [ρ]) as beginning either at the beginning of a word or with a stressed syllable and ending at the end of the word or before a stressed syllable. Once again, some cases of enclitics should be considered as forming part of the preceding word.

With this formulation, we can now combine the two levels of representation into a single prosodic tree as in (15):

(15)



Jassem's categories can be derived from this representation: an NRU is the leftmost [ρ] in a [τ], provided it does not begin with [^], and all other consecutive [ρ]s can be grouped to form Jassem's ANA.

It can be seen that the rhythm unit [ρ] and the tonal unit [τ], thus defined, form an interface between the lexical representation of the word and the actual pronunciation of the utterance, or between stress and accent as defined by Bolinger (1958).

In a study of the segmental duration of a large corpus of English, Hirst and Bouzon (2005) found that, as predicted by Jassem, but contrary to Halliday's model, word boundaries do play an important role in the rhythmic structure of English. Strong, negative correlations were found between the duration of a segment and the number of phonemes in the stress foot, in the NRU, and in the word, but no similar effect was found either in the syllable or in the ANA. Moreover the correlation was greater for the NRU than either the stress foot or the word, thus confirming Jassem's predictions.

3.4.3 Emphasis and Terminality

It seems, then, that the prosodic functions corresponding to stress/accent, centre, and boundary can be more adequately accounted for by an explicit model of prosodic structure such as I have just outlined.

The remaining two functions I mentioned, *emphasis* and *terminality*, do not look as if they can be accounted for by a difference in prosodic structure. One possibility, which I have tentatively suggested (Hirst 1983, 1998), is that these features could be accounted for as prosodic morphemes, whose function is similar in many ways to sentence particles, as found in many languages. Wakefield (2009, 2012) develops this idea more explicitly with a direct comparison between Cantonese sentence particles and English intonation patterns.

I will assume here, for the sake of argument, that English contains (at least) two abstract prosodic morphemes, which I will represent as [E] for an emphatic particle and [Q] for a question particle, because one of the major functions of [±TERMINAL] is expressing prosodically a distinction between a statement and a question. If we want to distinguish between the intonation of questions and that of unfinished statements, a debatable issue (as I discuss in Hirst 1998), we could suppose a third particle for that.

Question particles are well known in the description of many languages, both tonal and nontonal, for example, *ma* in Mandarin Chinese and *li* in Russian. Emphatic particles are also common. In Bambara (Mali; Bird Hutchinson, and Kante 1977), a sentence such as:

(16) Muso fila bɛ́ Sísé fɛ. (*Sissé has two wives*)

can be modified by the emphatic particle *de*, giving:

(17) Muso de fila bɛ́ Sísé fɛ. (*Sissé has two WIVES*)
(18) Muso fila de bɛ́ Sísé fɛ. (*Sissé has TWO wives*)

or

(19) Muso fila bɛ́ Sísé de fɛ. (*SISSE has two wives*)

where each time, the emphatic particle *de* immediately follows the emphasized word.

In section 3.8, I return to the question of how these abstract particles may be interpreted prosodically.

3.5 Phonetic Representation

3.5.1 Models of Fundamental Frequency

The search for an appropriate scale for measuring fundamental frequency has been one part of a systematic attempt, in particular by researchers from Holland ('t Hart, Collier, and Cohen 1990), to develop a model of the way in which pitch is perceived. This was done by stylizing raw fundamental frequency patterns as a sequence of straight lines, such that when the stylized frequency is used to resynthesize the utterance, the result is judged to be perceptually equivalent to the original intonation pattern.

Another approach to modeling pitch has been to attempt to model the way in which pitch is produced by speakers. In particular, work by Fujisaki (1991) and his colleagues has applied a model of pitch production to a large number of languages, including several tone languages, analyzing an intonation pattern as the superposition of three underlying components: a global baseline component, a sequence of phrasal components, and a sequence of shorter accent components. These three components are added in the logarithmic domain to produce a raw fundamental frequency curve.

A third approach has been to develop acoustic models that are neither directly models of speech perception nor of speech production but are compatible with both. This is the approach that I follow in this presentation.

3.5.2 Micromelodic Effects

Fitting a raw F0 curve with a mathematical model is not a simple, straightforward problem due to the fact that fundamental frequency curves are not always continuous: unvoiced portions of the utterance have no associated F0. Even when the curve is continuous, it is often not smooth, and this type of irregularity is hard to model simply.

If we look at a two-second extract of an utterance, corresponding to the words, “More news about the Reverend Sun Myung M(oon),”[3] we notice (see figure 3.1a) that the beginning and end of the F0 curve are in fact fairly continuous and smooth. The reason for this becomes obvious when we look at the phonemes associated with the curve as in figure 3.1b.

The smooth portion at the beginning of the extract corresponds to the phonemes /mɔːnjuː/, whereas that at the end of the extract corresponds to /ʌnmjʌŋm/. All of these phonemes are fully sonorant, either vowels, semivowels, or nasals. The discontinuity and irregularity of the F0 curve are due to the presence of obstruents in the utterance: stops and fricatives, which either interrupt the curve (for voiceless obstruents) or make it irregular (for voiced obstruents). The effect of these consonants has been called *micromelodic*, as distinct from the *macromelodic* characteristics of larger pitch movements associated with accents and intonation patterns. Micromelodic effects, then, are caused by the aerodynamic characteristics of the articulation of different phonemes (Di Cristo and Hirst 1986). Phonemes like vowels and sonorants, which hardly obstruct the airflow, have virtually no micromelodic effect, whereas stops and fricatives disturb or interrupt the flow.

The raw fundamental frequency curve then can be thought of as the interaction between two components: a *micromelodic* component, which is conditioned by the segmental nature of the individual speech sounds, and a *macromelodic* component, which corresponds to the underlying laryngeal gesture. This corresponds to the observation by Nooteboom (1997) that we do not perceive the observable discontinuities of raw pitch patterns unless they are longer than about two hundred milliseconds, as if

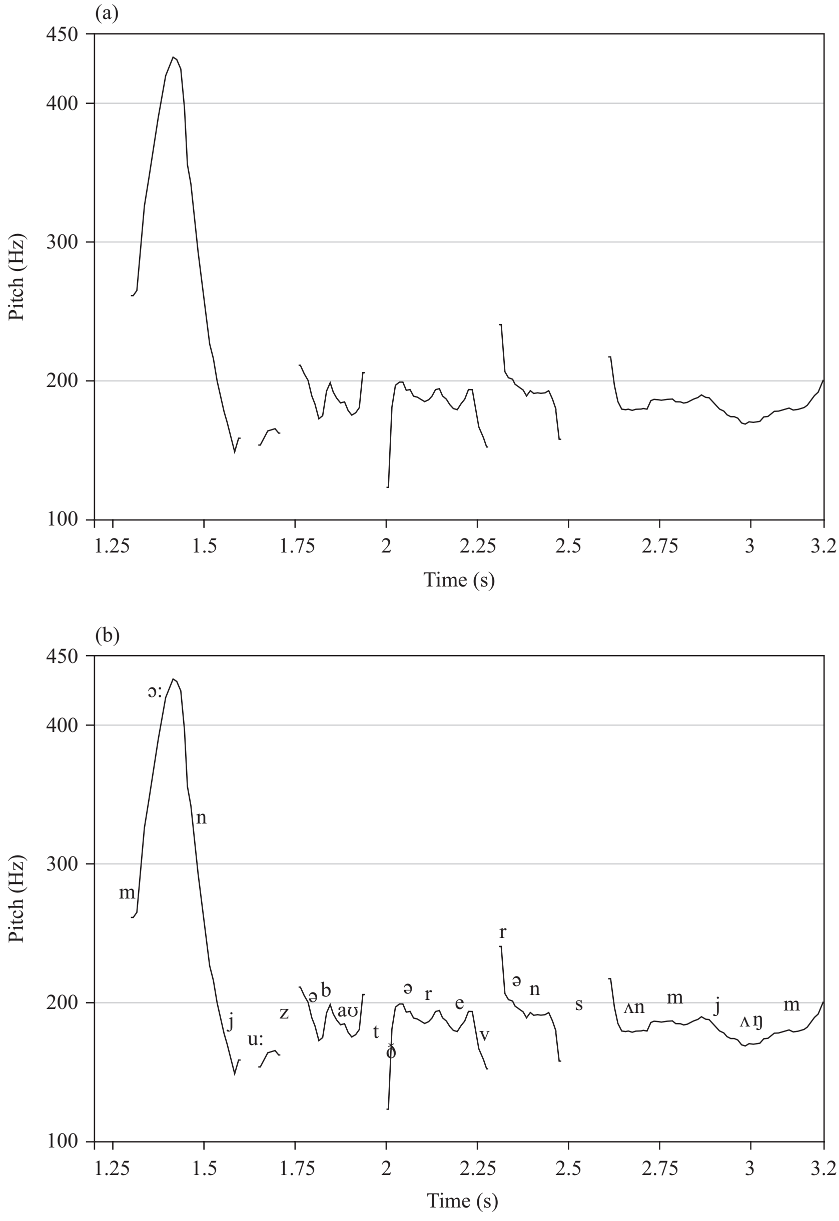


Figure 3.1
Two-second extract of the F0 curve corresponding to "More news about the Reverend Sun Myung M(oon)." (a) Raw F0. (b) F0 and phonemes.

human perception unconsciously bridges the silent gap by filling in the missing part of the pitch contour.

Linguists have been aware for a long time that fundamental frequency curves obtained from utterances containing only sonorants and vowels are much better behaved than raw F0 curves obtained from unrestricted speech. It is for this reason that linguists have often constructed sentences consisting of mainly sonorants and vowels, such as Eva Gårding's (1998) "Madame Marianne Mallarmé har en mandolin

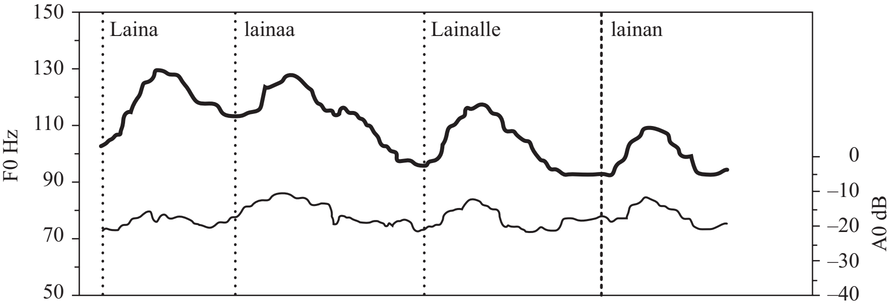


Figure 3.2
An example of a Finnish intonation pattern on a sentence with only sonorant phones. The sentence is "Laina lainaa Lainalla lainen" (Laina lends Laina a loan) (Iivonen 1998). The top curve is the fundamental frequency, and the lower curve is the intensity.

från Madrid" (Madam Marianne Mallarmé has a mandolin from Madrid) for Swedish; Hiroya Fujisaki's (2004) "Aoi aoinoewa yamanouenoieni aru" (The picture of the blue hollyhock is in a house on top of the hill) for Japanese; or Annti Iivonen's (1998) "Laina lainaa Lainalla lainen" (Laina lends Laina a loan) for Finnish (pictured in figure 3.2).

3.5.3 Macromelody and Micromelody

The idea, then, is that a raw intonation pattern is the interaction between two independent components: a macromelodic component determined by the accentuation and intonation of the utterance and a micromelodic component determined by the segmental phonemes.

If we compare two simple French phrases such as *À ton papa* (To your daddy) and *À ma maman.* (To my mummy), pronounced as statements, we can see that there is the same underlying macromelodic pattern for the two utterances and that the surface differences are simply due to the different phonemes of the utterances, voiceless stops in figure 3.3a and sonorant nasals in figure 3.3b.

What is particularly worth noting is that the F0 curve shown in figure 3.3a is practically superposable on that of figure 3.3b. It seems as if the F0 curve continues to change during the voiceless segments of the utterance, even though it is not visible. This is actually not surprising if we think in terms of a continuous change of tension of the vocal folds, which can continue to change even during voiceless segments.

If we now compare these patterns with those observed on the same phrases pronounced as questions, as in figure 3.4, we see once more that the two contours are practically superposable. And once again it seems clear that the F0 curve in figure 3.4a continues to change during the voiceless segments of the utterance.

In particular, the final rise on *papa* does not begin at the onset of the final vowel: the F0 at this point is already considerably higher than that at the end of the preceding vowel.

Notice that this idea of a continuously varying underlying pitch contour is not the model that is generally assumed in phonological descriptions of tonal and intonation contours. In the majority of these studies, it is assumed that tones are directly associated with vowels, which are the tone-bearing elements (see Goldsmith 1990, 44, for example) and that the fundamental frequency observed on the consonants is simply an interpolation between the tones on the vowels. If that were so, then it might be

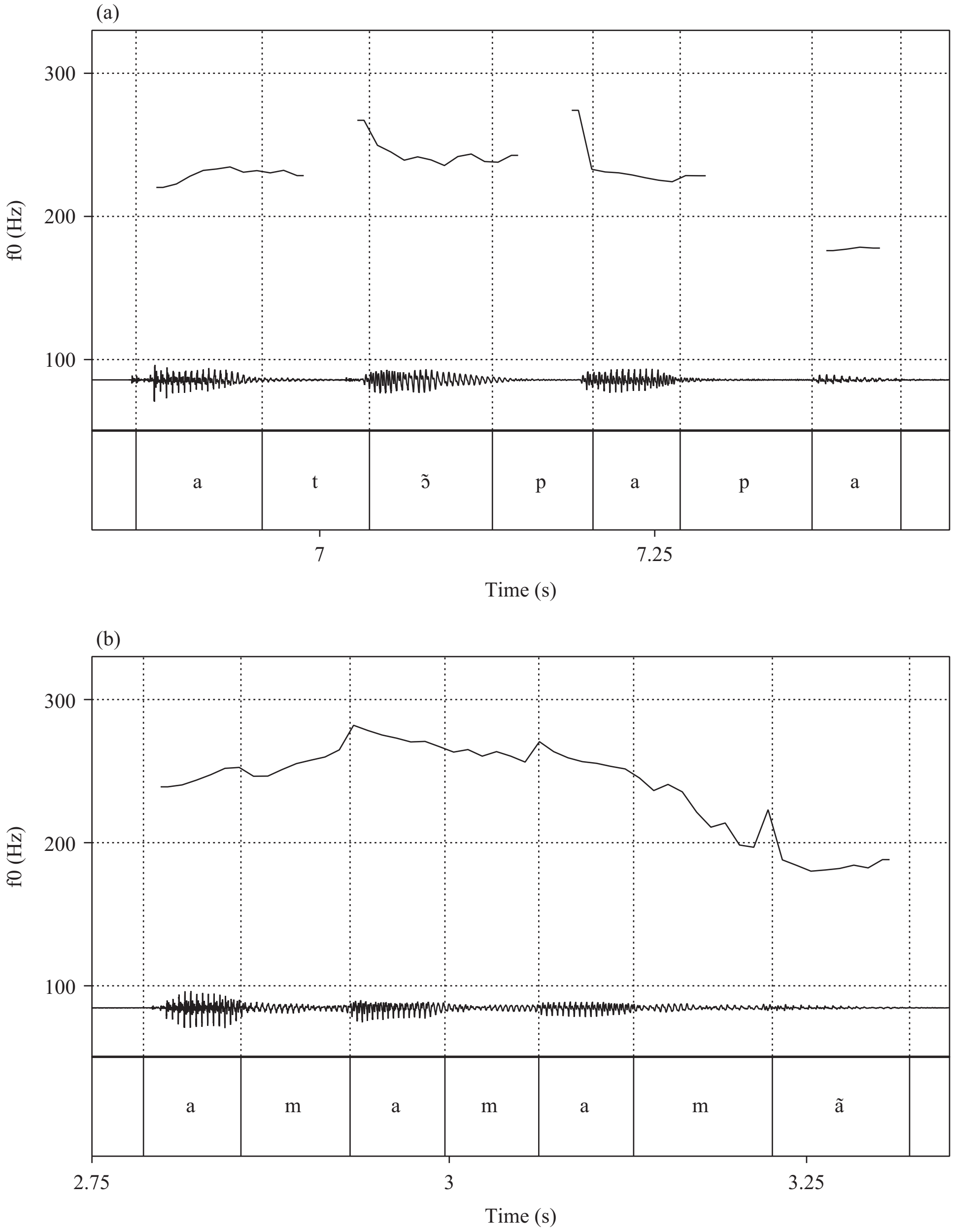


Figure 3.3
Two French phrases pronounced with a declarative intonation pattern. (a) *À ton papa*. (To your daddy.) (b) *À ma maman*. (To my mummy.)

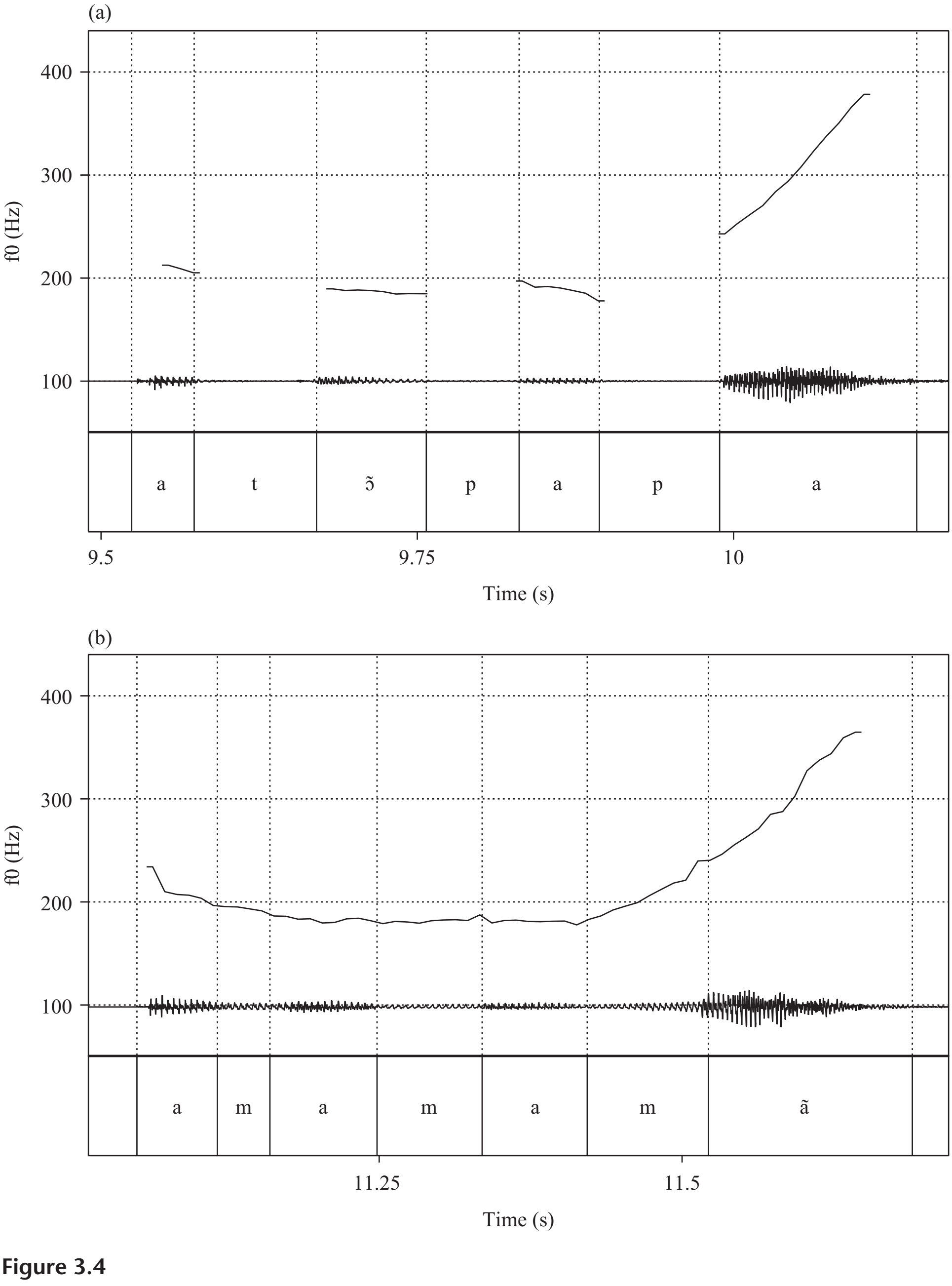


Figure 3.4
The same two French phrases as in figure 3.3 pronounced with an interrogative intonation pattern. (a) *À ton papa?* (To your daddy?). (b) *À ta maman?* (To my mummy?).

thought that the pitch curve visible in figures 3.3a and 3.4a is actually closer to the underlying form than that in figures 3.3b and 3.4b, which are simply the result of an interpolation on the sonorant consonants. The fact that the F0 curve follows the same trajectory in utterances with voiceless consonants as the smooth and continuous curve observed on the utterances with sonorants, however, and, in particular, the fact that the curve continues to evolve during the nonvoiced portions of the utterance, seems to me convincing evidence that the planning of these curves is the result of an underlying macromelodic pattern on which the micromelodic variations are subsequently superimposed.

3.5.4 A Model for F0 Curves

3.5.4.1 Macromelodic and micromelodic profiles The underlying macromelodic component of an intonation pattern, then, has, I shall assume, the two characteristics of being smooth and continuous. This is fortunate because modeling a discontinuous or irregular function is much more difficult than one that is continuous and smooth. To return to the example in figure 3.1, the underlying macromelodic profile of the curve might be something like the continuous curve in figure 3.5a, which could correspond to what we would produce if we were to hum the sentence instead of pronouncing it. I will return (section 3.5.5) to how the continuous curve was actually obtained; for the moment let us just assume that we have such a macromelodic profile.

Once we have a macromelodic profile, we can derive the micromelodic profile by dividing each value of the raw F0 curve by the corresponding value of the modeling function. This gives a result as displayed in figure 3.5b, where the smooth, continuous curve is the macromelodic profile and the discontinuous curve is the micromelodic profile derived as just described.

Notice that such a modeling technique is not simply a stylization of the raw F0 curve, because the raw curve has actually been factored into two orthogonal components without any loss of information. Multiplying each value of the continuous curve from figure 3.5b by that of the discontinuous curve in the same figure gives the original raw F0 as in the curve in figure 3.1a. For speech synthesis, it would be possible to model the micromelodic profile itself and to use this to improve the segmental quality of the utterance (for an application to Arabic, see Chentir, Guerti, and Hirst 2009). For the study of intonation, the resynthesis of the utterance with the macromelodic profile is generally of sufficiently high quality to evaluate the appropriateness of the modeled curve.

3.5.4.2 F0 transitions One of the simplest ways to model a smooth, continuous function such as that in figure 3.5a is as a piecewise sequence of transitions between successive points on the curve. In previous work, until recently, I referred to these points, following a fairly long tradition, as *target* points, but it should be noted that this name was not intended to imply that the "target points" necessarily have any actual psychological reality for the speaker and listener. To make it clear that these points are not intended to represent cognitive targets, like those, for example, in the model proposed in Xu and Wang (2001), I now prefer to use the term *anchor points*, which more clearly corresponds to their role in describing a pitch curve.

The advantage of a piecewise function over a global function is that each segment of the curve is defined locally by its own set of parameters, which means that a modification of one portion of the curve does not entail modifications throughout the rest of the curve. The simplest model would be a linear transition between two anchor points as in figure 3.6a, where the transition is defined by the function:

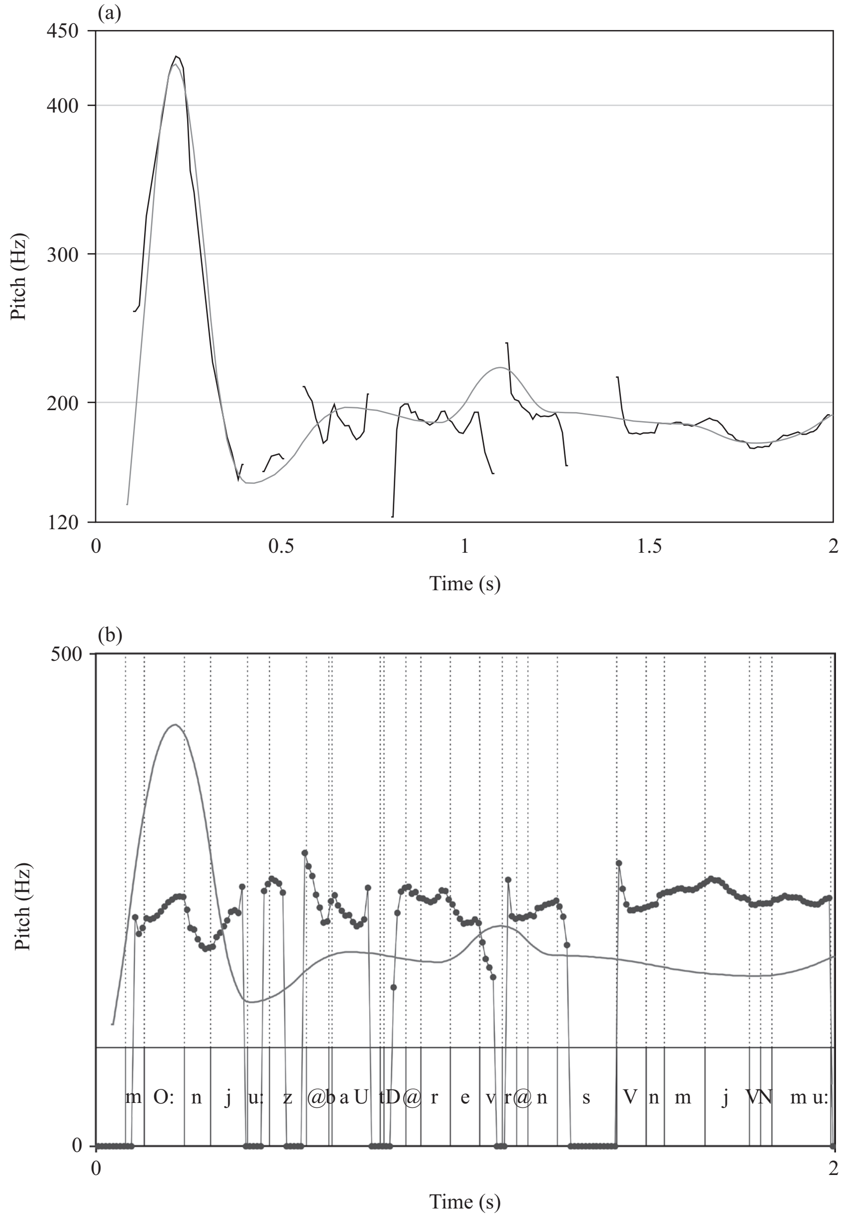


Figure 3.5
(a) Raw F0 (discontinuous) together with macromelodic profile (continuous) for the first two seconds of recording A01. (b) Same as panel (a), with the superimposed micromelodic profile.

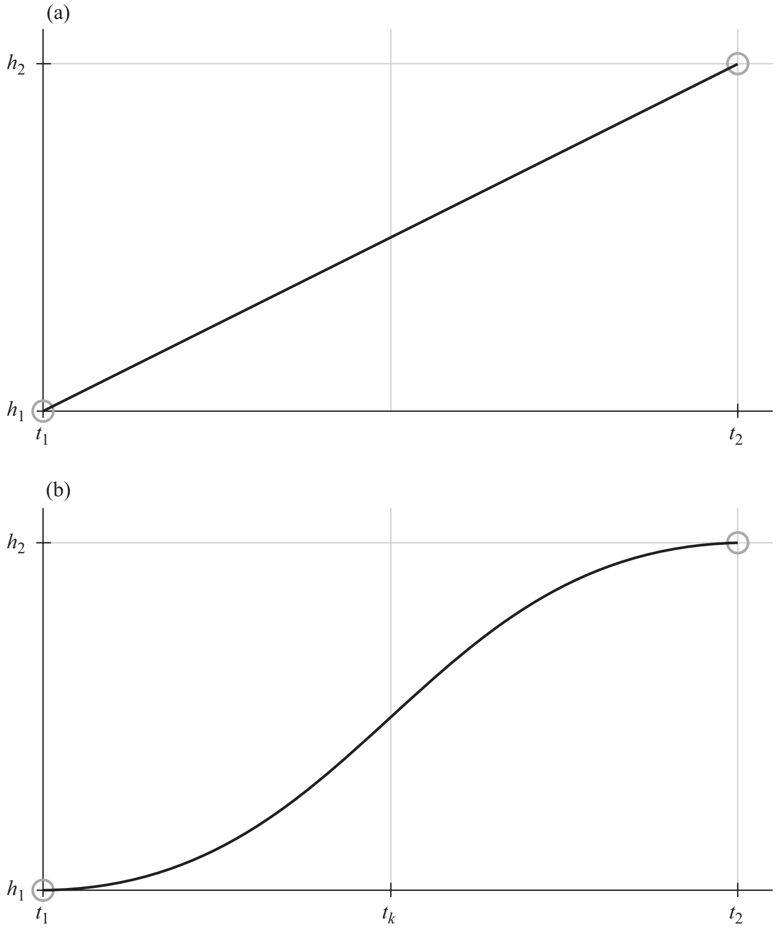


Figure 3.6
Linear and quadratic transitions from a first anchor point $<t_1, h_1>$ to a second point $<t_2, h_2>$. (a) A linear F0 transition. (b) A quadratic F0 transition.

$$h_i = h_1 + \frac{(t_2 - t_i)}{(t_2 - t_1)} \cdot (h_2 - h_1) \tag{3.1}$$

where h_1 and h_2 are the F0 values of two adjacent anchor points and t_1 and t_2 are the corresponding time values of these anchor points. Here the F0 value (h_2) of the second anchor point is higher than that (h_1) of the first anchor point, but the same reasoning would apply if it had been lower.

Naturally occurring F0 curves are curvilinear, not linear. A number of mathematical functions have been used in the past to model such functions. One of the simplest of these is a quadratic transition, corresponding to a constant acceleration followed by a constant deceleration of the pitch change as shown in figure 3.6b and as defined by the functions:

$$t_i \in [t_1 \ . \ . \ . \ t_k] : h_i = h_1 + \frac{(h_2 - h_1) \cdot (t_i - t_1)^2}{(t_k - t_1)(t_2 - t_1)} \tag{3.2}$$

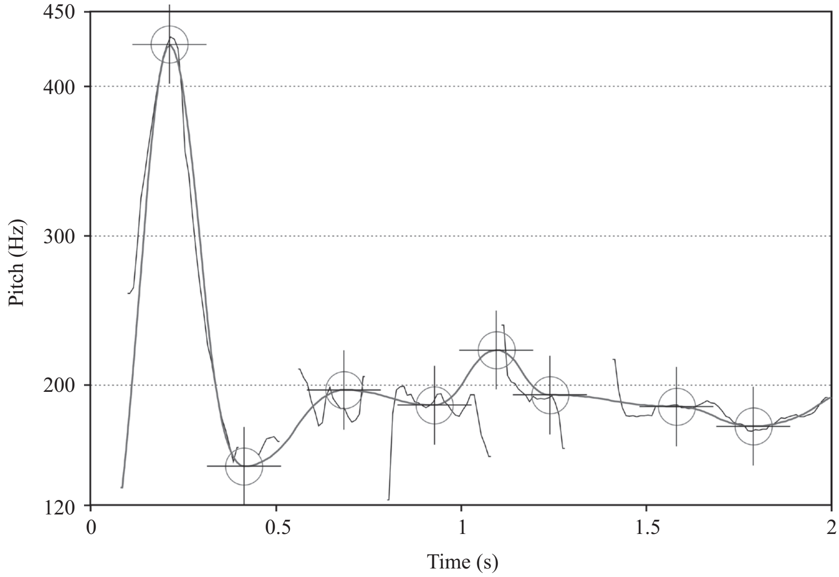


Figure 3.7
Macromelodic profile for a two-second extract from recording A01, defined as quadratic transitions between anchor points.

$$t_i \in [t_k \;...\; t_2] : h_i = h_2 + \frac{(h_1 - h_2) \cdot (t_i - t_2)^2}{(t_k - t_2)(t_1 - t_2)} \tag{3.3}$$

As can be seen from figure 3.6b, in the case of a rise, the transition consists of a concave curve from time t_1 to time t_k, the point of maximum slope, followed by a convex curve from t_k to t_2.

Figure 3.7 shows the same extract we have seen several times now, with the anchor points that define the curve represented as circles.

3.5.5 Momel

A piecewise quadratic function such as that illustrated here is known as a quadratic *spline* function and has been in use in our laboratory since the early 1980s to model intonation patterns using an algorithm called *Momel* (for “modeling melody”).

The Momel representation is formally equivalent to a subset of the contours that can be produced by the rise/fall/connection (RFC) model of intonation later developed by Paul Taylor (1994) as a tool for speech synthesis. The only difference is that the RFC model allows linear interpolations between two successive anchor points, as well as quadratic interpolations. If two successive anchor points have the same value of F0, then the transition will be linear (i.e., flat) with Momel too. I have personally never observed a case where a nonflat linear transition gives a better approximation to an F0 curve than a quadratic one.

The original implementation of Momel allowed the user to define anchor points manually by clicking on a representation of the F0 curve on the computer screen. These anchor points were converted into a quadratic spline function, and the user could then resynthesize the utterance using PSOLA resynthesis. Today, this can be done with Praat (Boersma and Weenink [1992] 2015) by creating a manipulation object and then removing and adding pitch points manually. Praat displays the pitch curve with linear interpolation between the pitch points, but an approximation of the quadratic spline function can be obtained by the command interpolate quadratically . . . , which was my own modest contribution to the Praat software.

Manual modeling of F0 is highly subjective, and it was for this reason that my colleague Robert Espesser and I developed an automatic version of the algorithm (Hirst and Espesser 1993) based on our experience of using the manual implementation of the model over a period of several years. The algorithm, which is described in detail in Hirst, Di Cristo, and Espesser (2000), uses a form of robust regression to optimize the modeling of raw fundamental frequency curves with a quadratic spline function.

The algorithm was later evaluated on a corpus of read speech in five languages (Eurom1 corpus) during the course of the Multext European project (Véronis, Hirst, and Ide 1994). Evaluators were instructed to add or delete anchor points of the modeled speech only when such corrections made an audible improvement to the resynthesis of the speech. Table 3.1 shows the statistics[4] for these corrections for the corpus of read speech, together with those for a corpus of spontaneous speech in French.

The results of the evaluation were very encouraging. The F-measures for the different languages showed a global efficiency of around 95 percent, and even the corpus of spontaneous French showed an F-measure of 93.4 percent even though the algorithm had not at all been optimized for spontaneous speech.

Examination of the errors in the anchor points showed that one type of error in particular occurred systematically. This concerned a pitch rise before a silent pause where, frequently, the algorithm missed the final rise entirely. An example of this type of error can be seen in figure 3.8. This error is understandable because the algorithm uses a local modeling technique, fitting a parabola to portions of the curve. As can be seen in the example, the raw F0 exhibits a concave portion corresponding to the acceleration of the pitch rise but hardly any convex portion corresponding to the decelerating portion of the rise. This is the reason why the original algorithm fails to produce a final anchor point.

The Momel algorithm has since been implemented as a Praat plug-in (Hirst 2007), which allows users to use its functions directly from the Praat menus without needing

Table 3.1

Results for the evaluation of the automatic Momel algorithm on read speech for English, French, German, Italian, and Spanish and for spontaneous speech in French (Fref corpus)

		Number of points			Evaluation		
Corpus	Language	Detected	Added	Deleted	Recall	Precision	F-measure
Eurom	English	8,380	623	125	93.0	98.5	95.7
	French	6,547	423	130	93.8	98.0	95.9
	German	13,595	1,145	506	92.0	96.3	94.1
	Italian	9,475	337	330	96.4	96.5	96.5
	Spanish	8,985	651	16	93.2	99.8	96.4
Fref	French	9,835	532	744	94.5	92.4	93.4

Note: The columns show the total number of anchor points detected, number of anchor points added manually, number of anchor points deleted, and the statistical measures of recall, precision and F-measure. *Source*: Data from Campione (2001).

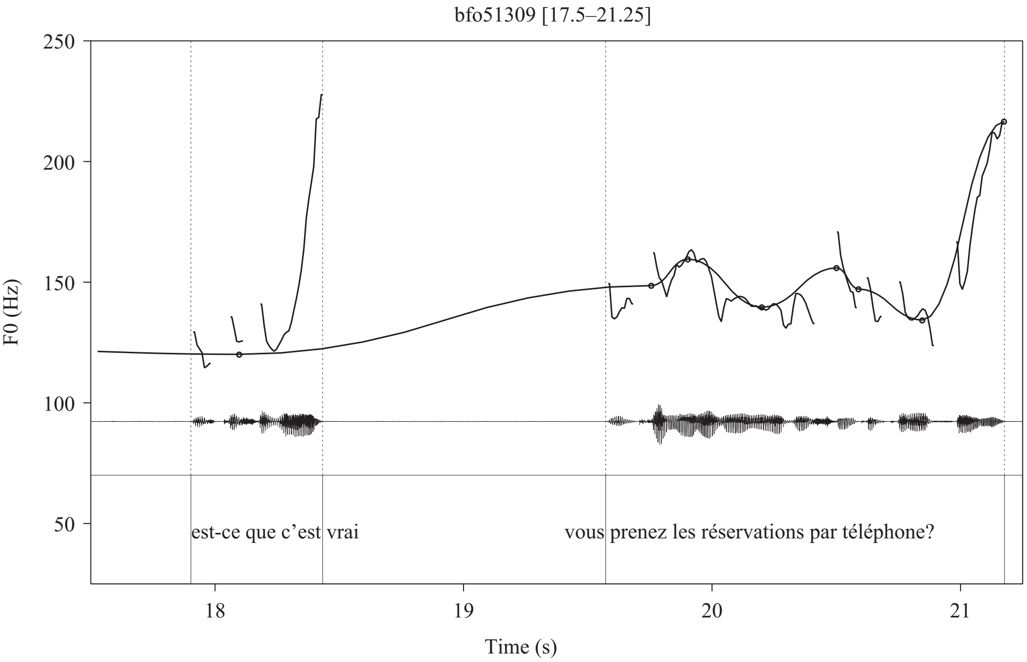


Figure 3.8
Old version of the automatic Momel algorithm for the French utterance, *Est-ce que c'est vrai? Vous prenez les réservations par téléphone?* (Is it true? You take bookings by phone?). Raw (discontinuous) and modeled F0 (continuous).

to handle scripts. The systematic error we have just seen was corrected by a special treatment before silent pauses. The concave part of the pitch rise is now extended to include a similar-shaped convex portion. In other words, to obtain the best fit for this pitch rise, a high anchor point is calculated and situated in the silent pause as near as possible to the previous low anchor point. The anchor point is calculated so that the concave portion of the pitch rise follows the raw data points as closely as possible. The result of this improved algorithm can be seen in figure 3.9.

An evaluation of the improved algorithm was carried out on a corpus of read speech in Korean (Hirst et al. 2007). It showed a significant and systematic improvement as compared to the older version of the algorithm. It is, naturally, desirable that the modeling tools we use should be as theory-neutral as possible. Complete neutrality, though, is obviously not entirely feasible, because any model necessarily makes some assumptions about the nature of underlying representations, as we saw in the discussion (section 3.5.4.1) of whether the underlying contour should be based only on the contours observed on the vowels or whether it should be modeled as a continuous underlying contour.

Rather than suggest that Momel is theory-neutral, then, I like to think that it is what we could call *theory-friendly*. I believe that the algorithm can be compatible with a number of different theoretical approaches to the description of speech melody. It has been used in the past as a first step for modeling with the Fujisaki model (Mixdorff 1999). It has also been used as first step for ToBI for both English (Maghbouleh 1998;

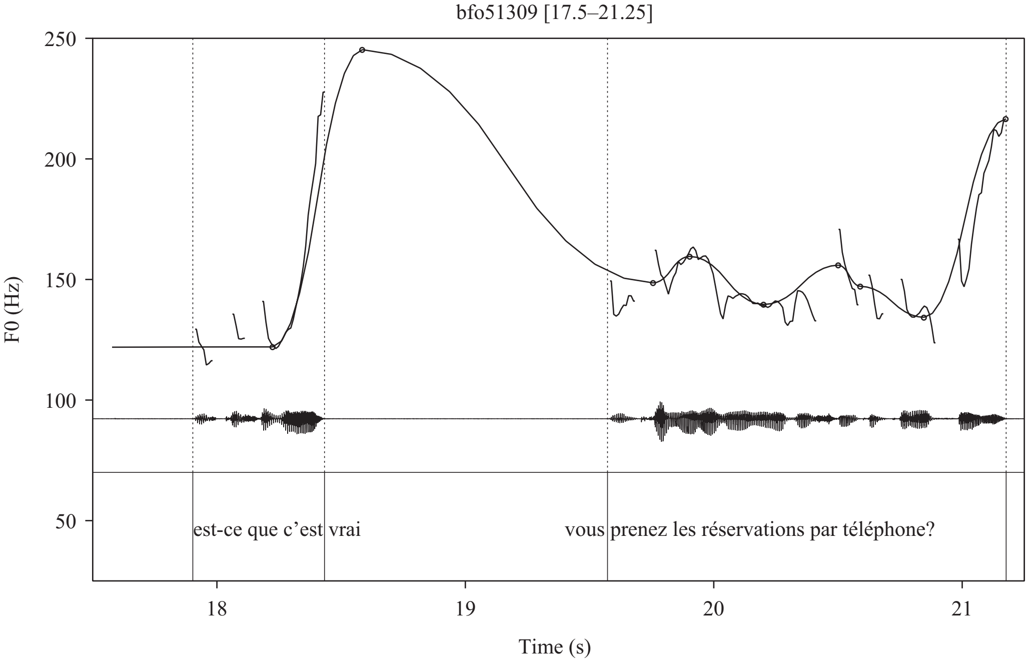


Figure 3.9
Improved version of the automatic Momel algorithm for the same utterance as in figure 3.8. Raw (discontinuous) and modeled F0 (continuous).

Wightman and Campbell 1995) and Korean (K-ToBI) (Cho and Rauzy 2008). It is also compatible with our own surface phonological representation alphabet, INTSINT.

3.6 Surface Phonological Representation

3.6.1 INTSINT: An International Transcription System for Intonation

Momel, as we have seen, provides a reversible modeling of the raw F0 with no loss of information because the raw F0 is factored into two components, macromelody and micromelody. Multiplying the two together recovers the original F0.

INTSINT was originally developed as a tool for linguists to provide a surface phonological representation of an intonation pattern. The original version of the system (Hirst 1987) was based on an inventory of minimal pitch contrasts found in published descriptions of the intonation patterns of numerous languages. The aim was to provide a tool for the systematic description of these intonation patterns, something along the lines of a narrow transcription using the IPA. Like the IPA, it was intended that INTSINT could be used for preliminary descriptions of intonation patterns, even for languages that had not previously been described. Notice that this aim is very different from that of the ToBI system (Silverman et al. 1992), which presupposes that the inventory of intonation patterns for the language being described has already been established. The official website for ToBI (http://www.ling.ohio-state.edu/tobi/) makes this particularly explicit:

> Note: ToBI is not an International Phonetic Alphabet for prosody. Because intonation and prosodic organization differ from language to language, and often from dialect to

dialect within a language, there are many different ToBI systems, each one specific to a language variety and the community of researchers working on that language variety.

The INTSINT system (whose name was suggested to us by Hans 't Hart in a personal communication) was presented in Hirst and Di Cristo (1998) and was used there for the annotation of the intonation of nine different languages. Basically, it describes an intonation contour as a sequence of tonal segments labeled using an alphabet of eight symbols. The tonal segments are assumed to be of three types:

Absolute tones: *t* (top), *m* (mid), *b* (bottom). These are assumed to refer to the corresponding position of the speaker's current pitch range.
Relative tones: *h* (higher), *s* (same), *l* (lower). Unlike absolute tones, relative tones are assumed to be refined with respect to the preceding tonal segment.
Iterative relative tones: *u* (upstepped), *d* (downstepped). These are also defined relative to the preceding tonal segment but generally involve smaller pitch changes and often occur in a sequence of steps either upward or downward.

In the chapters in Hirst and Di Cristo, the INTSINT tones were represented by graphic symbols between two horizontal lines and aligned with the text. These symbols were top [⇑], bottom [⇓], higher [↑], same [→], lower [↓], upstepped [<], and downstepped [>]. The mid tone was reserved for the unmarked onset of an intonation unit and was not marked.

In most later publications, the capital letters T, M, B, H, S, L, U, and D were used instead of the graphic symbols. Since Hirst (2011), I prefer to represent the INTSINT tones with lowercase rather than uppercase letters. This may help to avoid confusion with other more abstract coding schemes such as ToBI (Silverman et al. 1992), or the even more abstract underlying representation used in Hirst (1998) and section 3.8, both of which use some of the same letters as INTSINT.

3.6.2 Mapping from INTSINT to Momel

Although INTSINT was introduced as a descriptive tool for linguists, since its introduction was later than the creation of the F0 modeling tool, I already had in mind the possibility that this surface phonological annotation could be linked to the analysis of the F0 curve as a sequence of phonetic anchor points. I anticipated, then, that it might be possible to map the output of the Momel algorithm onto a sequence of symbols from the INTSINT alphabet. To do this, following the idea of an analysis by synthesis paradigm, it was first necessary to define a mapping in the other direction, that of synthesis. Some of the history of the way in which this mapping was defined is described in Hirst (2005).

In its current implementation, the mapping depends on two speaker-/utterance-specific parameters called *key* and *span*, which together define the speaker's pitch range.[5] The *key* (like a musical key) defines a central reference point for the speaker's pitch range, and the *span* defines the maximum and minimum pitch values of the range, which are taken to be symmetrical (on a logarithmic scale) above and below the speaker's key.

The two parameters together define three absolute tones—top, mid, and bottom—with respect to the speaker's pitch range, as in the following formulas, which assume that the value of key is given in hertz and the value of span in octaves:

$$t = key \times \sqrt{2^{span}} \quad m = key \quad b = \frac{key}{\sqrt{2^{span}}} \tag{3.4}$$

The pitch anchor points corresponding to the relative tones are then defined with respect to both the preceding anchor point (here called *p*) and the top (*t*) or bottom (*b*) of the range.

An anchor point coded *h* is simply defined as the geometric mean (i.e., the mean on a log scale) of the preceding anchor point and the top of the range. It thus corresponds to a pitch movement that moves up halfway toward the value of *t*. As can be expected, an anchor point coded *s* is defined as the same as the preceding anchor point. Symmetrically to *h*, an anchor point coded *l* is defined as the (geometric) mean of the preceding anchor point and the bottom of the range.

$$h=\sqrt{p*t};\quad s=p;\quad l=\sqrt{p*b} \tag{3.5}$$

For the iterative tonal segments *u* and *d*, the implementation defines the values as the (geometric) mean of the value of the preceding anchor point and that which would be obtained if the anchor point were coded *h* or *l*, respectively. In other words, these anchor points correspond to a pitch excursion one-quarter of the way to the top or bottom of the pitch range.

$$u=\sqrt{p*\sqrt{p*t}};\quad l=\sqrt{p*\sqrt{p*b}} \tag{3.6}$$

Assuming, once again, a logarithmic scale for the pitch range, these values are illustrated graphically in figure 3.10.

This implementation obviously makes a number of assumptions, most of which would be open to empirical investigation. This is one of the major advantages of an explicit model such as this.

One consequence of the model, which was not specifically intended but which turns out to be fortunate, is that a sequence of alternating *h* and *l* tones will automatically introduce an iterative lowering of the tones, much like the probably universal effect of downdrift that has been described in the literature on tone and intonation as occurring in languages throughout the world. This downdrift effect, as illustrated in figure 3.11, is thus an automatic by-product of the way in which the relative tones are defined.

Since it was first developed, the Momel algorithm has been applied relatively successfully to a number of different languages, including English, French, Italian, Catalan,

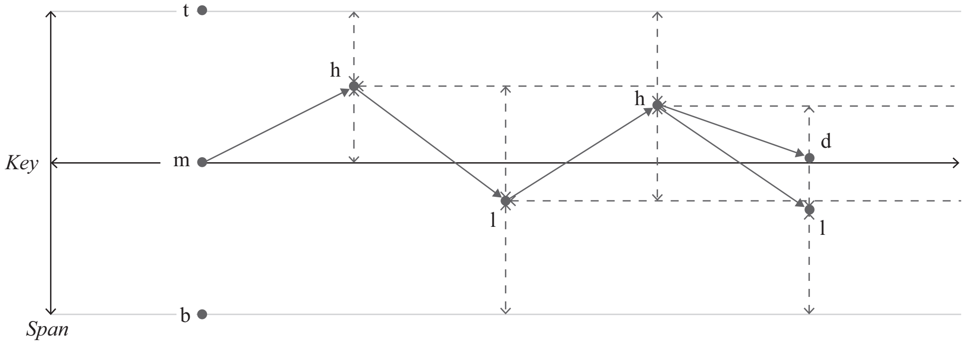


Figure 3.10
Graphic illustration of the mapping from INTSINT to Momel with pitch range defined by the two parameters key and span.

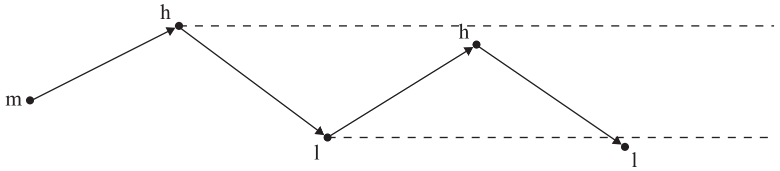


Figure 3.11
A graphic illustration of the fact that in this implementation, downdrift is an automatic by-product of the way in which the relative tones are defined.

Brazilian Portuguese, Venezuelan Spanish, Russian, Arabic, isiZulu, and Korean (for references, see Hirst 2007). More recently (Zhi, Hirst, and Bertinetto 2010), the algorithm was applied to a corpus of speech in Standard (Beijing) Chinese. This was particularly challenging, because the corpus used was spontaneous speech and involved a language with a rich lexical tone system. An attempt to optimize window size for the algorithm showed no overall improvement with respect to the manually corrected data. This was taken to confirm the fact (as suggested by Xu and Sun 2002) that pitch change in a lexical tone language like Chinese is not notably faster than in languages with no lexical tone. The annotated data obtained during this application will constitute a useful yardstick for evaluating improvements to the automatic algorithm, which is expected to be far more robust than data annotated for languages with no lexical tone.

3.6.3 Mapping from Momel to INTSINT

Having defined a mapping from tonal segments (INTSINT) to pitch anchor points (Momel), the same model can be used to establish a reverse mapping from the anchor points to the tonal segments. As is generally the case, such a reverse mapping from continuous variables to discrete categories is much less straightforward than the mapping from categories to continuous variables. The approach we have adopted is an exhaustive search of the target space for the optimal values of the two parameters key and span, together with the optimal coding of the sequence of anchor points, given those two parameters.

The procedure, as described in Hirst (2005), has been implemented as a Perl script. It assumes that the relevant target space is defined as follows:

$$\begin{aligned}&\text{key} = \text{mean} \pm 50 \text{ Hz (step: 1)}\\&\text{span} = 0.5 \ldots 2.5 \text{ octaves (step: 0.1)}\end{aligned} \tag{3.7}$$

The script thus tries each of the possible values of the two parameters within this target space. For each of the two thousand possible couples ⟨*key;span*⟩ the script evaluates every possible coding of the anchor points using the formulas in equations (3.4), (3.5), and (3.6) and calculating the sum of the square of differences between the predicted value and the observed value. The output of the script is thus the optimized value (within the target range) of the parameters *key* and *span*, together with the optimal INTSINT coding using these parameters. The output of the script is a text file such as in (20), corresponding to the application of the script to the same extract from recording A01 that was shown in figure 3.5 above and elsewhere.

(20)

```
; A01_01.intsint created on Tue Aug 24 08:12:47 2010
; by intsint.pl 2.11 ; from A01_01.momel
;    32 values       mean = 191 <parameter span = 1.4> <parameter key = 235>
0.113     M     221     235
0.219     D     205     208
0.434     D     182     190
0.746     B     120     145
1.177     S     120     145
1.423     T     428     382
1.623     B     146     145
1.894     U     197     184
. . .
```

In this output, the optimized values of key and span (here, 235 Hz and 1.4 octaves, respectively) are given, together with the sequence of anchor points. After the values for key and span, each line gives the time value (in seconds), the optimized INTSINT code, the original anchor point used as input, and the predicted anchor point derived from the coding with the current values of key and span.

Unlike with Momel, there is some loss of information with the INTSINT coding (as can be seen from the differences between the third and fourth columns of the output in (20); the model is still, however, a reversible one that can be used for synthesis. Figure 3.12 shows the result of recalculating the pitch anchor points from the results of the automatic INTSINT analysis, applied to a five-sentence passage from the Eurom1 corpus, and compared to the original anchor points obtained from the application of the Momel algorithm.

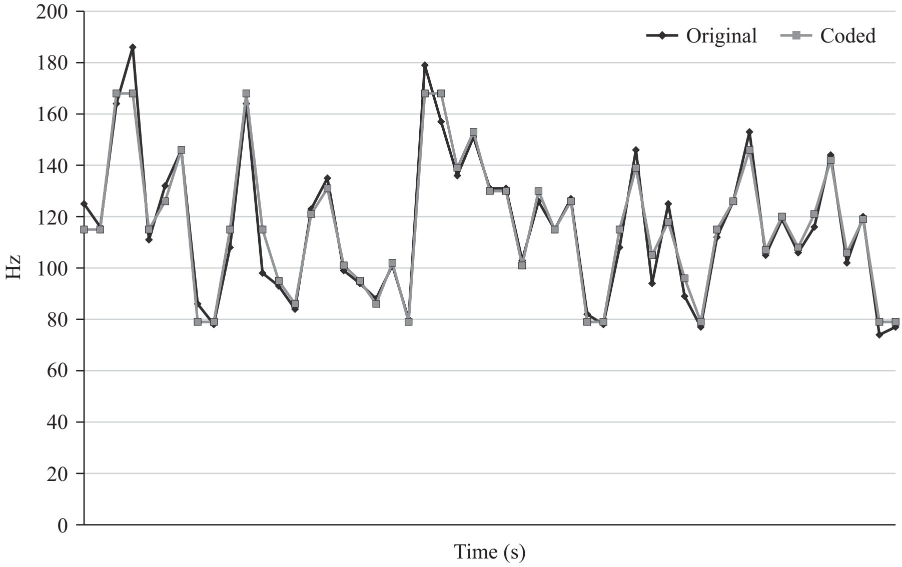


Figure 3.12
Coded (grey squares) versus original values (black diamonds) of pitch anchor points for a five-sentence passage after automatic coding with the INTSINT alphabet.

3.6.4 Longer-Term Characteristics of Pitch Range

The implementation of INTSINT as I have described presupposes that there are no variations in key and span within the segment of speech that is analyzed. In authentic speech, such changes naturally occur quite frequently and are obviously very significant and important. One solution is to implement the INTSINT coding on smaller segments of speech such as breath groups, making what seems a fairly reasonable assumption that changes of key and/or span are more likely to occur between breath groups rather than within them. For an investigation of the possibility of automatizing such a process, see De Looze (2010), which implements an algorithm (AdoReVa) applying a cluster analysis for this task.

3.7 ProZed

The availability of explicit models of speech melody such as those I have described in this chapter makes it possible to use such models to implement more abstract representations of prosody. With this aim in mind, my current work in progress concerns the implementation of a prosody editor for linguists called *ProZed* (Hirst 2015), which aims to provide linguists with a tool allowing them to experiment with different abstract phonological models, providing them with an acoustic output with which they can, at least informally, evaluate the relative value of different models.

A version of this editor applied to speech rhythm is described in Hirst and Auran (2005), which implemented an empirical linear model for rhythm where each rhythm unit is characterized by three parameters: q, a (possibly speaker-dependent) parameter defining a unit of quantal lengthening; t, a long-term parameter of tempo; and k, a local scalar effect of lengthening specific to each rhythm unit. With these parameters the duration of the rhythm unit is defined as:

$$\hat{d}_\rho = t \cdot \left\{ \sum_{i=1}^{m} \bar{d}_{i/p} + k \cdot q \right\} \tag{3.8}$$

where di/p is the mean duration of all the phones labeled as the same phoneme p as that occurring in position i in the rhythm unit.

ProZed has been implemented as a plug-in to the Praat software. It allows the manipulation of the rhythmic and tonal aspects of speech as defined on three specific tiers, in addition to the phoneme tier. The first two of these are the RU (rhythm unit) tier and the TU (tonal unit) tier. These two tiers control the short-term variability of segmental timing and tonal variation, respectively. Longer-term variations of rhythm and melody can be controlled via a third tier, the IU (intonation unit) tier.

The speech input to the program may be natural recorded speech, the prosodic characteristics of which will then be modified by the software, or, alternatively, it may be the output of a speech synthesis system with, for example, fixed durations for each speech segment.

The current version of the program is designed as the resynthesis step of what is planned to be a complete analysis by synthesis cycle. This will be directly integrated with the output of the Momel-INTSINT and ProZed rhythm analysis models and interfaced with the automatic alignment system SPPAS (Speech Phonetization Alignment and Syllabification; Bigi and Hirst 2012, 2013).

3.8 Underlying Phonological Representation

We have seen that it is possible to derive automatically from the acoustic signal a phonetic representation of an intonation contour (Momel) and that this can then be converted automatically into a time-aligned sequence of discrete symbols (INTSINT), which, together with the output of the automatic alignment software, can constitute at least a first approximation to a surface phonological representation of the prosody of an utterance. It now remains to be seen how we can bridge the gap between the abstract functional representation of prosody outlined in section 3.3 and this surface phonological representation.

For British English, the classical description of the intonation of nonemphatic declarative utterances and *wh*-questions, as in Armstrong and Ward (1926), Kingdon (1958), and O'Connor and Arnold (1961), is of a pitch pattern that begins on a mid-level with a rise to a high pitch on the first accented syllable, followed by a step-like lowering of pitch on each accented syllable until the last accented syllable, followed by a fall to the bottom of the speaker's pitch range.

Among the examples of this "tune," Armstong and Ward (1926) give:

(21) It was the ‘last ‘thing I ex- ‘pected to ‘find there.

In terms of prosodic structure, as described in section 3.4, this utterance can be parsed into one intonation unit containing six tonal units:

(22) [[^ it was the] [last] [thing I ex-] [-pected to] [find there]]

The INTSINT annotation of the tonal representation of this utterance (ignoring here the question of the relative alignment of the tones and the segmental material) could then be:

(23) [**m** It was the [**h** last] [**d** thing I ex-] [**d** -pected to] [**d** find there **b**] **b**]

where the initial [m] and final [b] tones are attached directly to the intonation unit and the other tones are attached to the tonal units.

Pike (1945) describes a neutral intonation pattern for declarative utterances in American English as a sequence of falling contours associated with each accented syllable. This could be annotated as:

(24) [**m** it was the [**h** last **l**] [**h** thing I ex- **l**] [**h** -pected to **l**] [**h** find there **b**] **b**]

Pike characterizes a "descending stress series" such as in (23) as expressing "EXTREME precision, or certainty" (70). Because of this difference in interpretation, to American ears the British intonation may sound overprecise or pompous, while to British ears the American pattern may sound overenthusiastic or exuberant.

This dialectal difference between American and British English intonation patterns is interesting since a downstepped tone in many African tone languages can often be traced to a historical low tone, which is sometimes described as a *floating* tone (Clements and Ford 1979), that is, a tone that is not phonetically realized as a pitch target but has the effect of lowering the following high tone.

The two dialects could consequently be described as having the same underlying representation, with the difference that in British English, the low tone is somehow "delinked" from the prosodic structure.

We can compare this intonation to that of a French translation of (22):

(25) C'était la 'dernière 'chose que je m'atten-'dais à 'trouver 'là.

We are likely to find that instead of a sequence of falling pitch movements as in American English (as described by Pike 1945) or a sequence of downsteps as in British English, the French utterance is produced as a sequence of rising pitch movements, culminating with the accented syllable and with a final falling pitch on the last syllable (Hirst and Di Cristo 1984; Di Cristo 1998).

We have suggested (Hirst and Di Cristo 1984) that one of the major prosodic differences between English and French is that in English, tonal units consist of an accented syllable *followed* by a number of unaccented syllables, and in French, they consist of an accented syllable *preceded* by a number of unaccented syllables. Example (25) could consequently be annotated in the following way:[6]

(26) [[c'était la der-] [-nière chose] [que je m'attendais] [à trou-] [-ver là]].

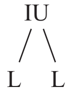
with which we can associate an INTSINT representation:

(27) [**m** [**m** c'était la der- **h**] [**l** -nière chose **h**] [**l** que je m'attendais **h**] [**l** à trou- **h**]
[**d** -ver là] **b**]

For an underlying phonological representation of these tunes, I will assume, following Pierrehumbert (1980), that there are only two values for underlying tones, L and H, and that the specific INTSINT interpretations of these underlying tones are allophonic variants determined by the context. Under this interpretation, it seems clear, then, that for all three varieties we have looked at—American English, British English, and French—the underlying tones associated directly with the intonation unit (= boundary tones) are [L L].

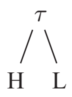
We can assume, then, a template like the following to account for this association:

(28)



For the tonal units, we can assume that both British English patterns and American English patterns are derived from a single template:

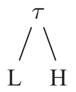
(29)



and that in British English, but not in American English, there is a downstepping rule that will have the effect of interpreting an underlying sequence [H L] [H . . .] as [h] [d . . .]

French patterns can be derived from:

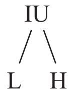
(30)



assuming simply that in French, a downstepping rule applies only to the last tonal unit, interpreting [. . . H] [L H] as [. . . h] [d].

For the intonation of yes/no questions in English and French, the only change that needs to be made to the underlying representation for statements is that the final L is replaced by H, triggered by the presence of the question particle [Q] so that we have:

(31)



For French, the presence of a question particle triggers not only the application of the template in (31) but also the application of a downstepping rule on each accent so that a sequence [. . . H] [L H] is interpreted as [. . . h] [d] for each tonal unit in the intonation unit.

For English, the presence of an emphatic particle [E] has two effects. The pitch accents on the accented syllables are either deleted entirely or considerably reduced, as if there were an intermediate prosodic constituent between the tonal unit and the intonation unit:

(32) a. [[John has] [two] [wives]]. (*John has two wives*)
b. [[[John has] [two] [wives]]] (*JOHN has two wives*)
c. [[John has] [[two] [wives]]] (*John has TWO wives*)
d. [[John has] [two] [[wives]]] (*John has two WIVES*)

There might, then, be a condition that only the first tonal unit within this emphatic constituent is assigned tones. This cannot be the whole story, though, because the last example of (32) would then be assigned the same tonal representation as the first, unemphatic, example. In fact, the pitch accent on an emphatic constituent is systematically higher than other accents. One possibility, then, would be to assume that the emphatic constituent is assigned [H L] tones just like the tonal units, so the underlying representation of (32) would then be something like:

(33) a. [**L** [**H** John has **L**] [**H** two **L**] [**H** wives **L**] **L**] (=John has two wives)
b. [**L** [**H** [**H** John has **L**] [two] [wives] **L**] **L**] (=JOHN has two wives)
c. [**L** [**H** John has **L**] [**H** [**H** two **L**] [wives] **L**] **L**] (=John has TWO wives)
d. [**L** [**H** John has **L**] [**H** two **L**] [**H** [**H** wives **L**] **L**] **L**] (=John has two WIVES)

which could then be interpreted as the INTSINT annotations (for British English), where the **h** or **d** followed immediately by **h** would ensure a higher pitch on the following syllable than when there is just a single **h**:

(34) a. [**m** [**h** John has] [**d** two] [**d** wives **b**] **b**] (=John has two wives)
b. [**m** [**h** [**h** John has **b**] [two] [wives] **b**] **b**] (=JOHN has two wives)
c. [**m** [**h** John has] [**d** [**h** two **b**] [wives] **b**] **b**] (=John has TWO wives)
d. [**m** [**h** John has] [**d** two] [**d** [**h** wives **b**] **b**] **b**] (=John has two WIVES)

3.9 Conclusion

I have presented in this paper a sketch of a fairly complete model of prosody, although a great number of details concerning the implementation of the model need to be specified. The model extends from a functional annotation of intonation at the most abstract level, via an underlying phonological representation and a surface phonological representation, to a phonetic representation, which is directly convertible into an acoustic signal. The model is described in more detail in my book (Hirst 2020), together with more detailed justification of the approach presented here. It is obvious that at each level, I have made a number of arbitrary choices, which may not stand up to further investigation, but I am nonetheless convinced that a multilevel, multilingual model of this type is what is needed to further our knowledge of the complex way in which prosody contributes to our interpretation of utterances.

Notes

1. Faure (1962) makes a similar distinction between “les caractères” (forms) and “le rôle” (functions).

2. I would today prefer to call this [±ACCENT] rather than [±STRESS].

3. From the first recording (A01) of the Aix-Marsec corpus (Auran, Bouzon, and Hirst 2004).

4. The statistics used here were *recall* and *precision*, which are commonly used in information retrieval and binary classification tasks. *Recall* is the proportion of positive values that are correctly identified, and *precision* is the proportion of values identified as positive that are correct. In our context, precision is the percentage of all automatically detected targets that were considered correct, while recall is the percentage of all “correct” targets that were automatically detected. The *F-measure*, considered a measure of global efficiency, is a combination of the two values and is calculated as the harmonic mean of recall and precision (van Rijsbergen 1979).

5. In some earlier publications, I have used the word *range* for what I here call *span*. I now prefer to use *span* to refer to the interval, independent from the value of the *key*. The two values *key* and *span* together define the speaker’s *range*. So we might say that a given speaker has a pitch range from 100 Hz to 200 Hz, corresponding to a span of one octave and a key of 141 Hz.

6. French words are regularly accented on the final syllable, but nonfinal words may also be accented on the first syllable, especially when immediately followed by an accented syllable as in *dernière chose* and *trouver là* in this example. See Hirst and Di Cristo (1984) and Di Cristo (1998) for details.

References

Abercrombie, D. 1964. “Syllable Quantity and Enclitics in English.” In *In Honour of Daniel Jones*, edited by D. Abercrombie, D. Fry, P. MacCarthy, N. Scott, and J. Trim, 216–222. London: Longman.

Armstrong, L., and I. Ward. 1926. *A Handbook of English Intonation*. Cambridge: Heffer.

Auran, C., C. Bouzon, and D. J. Hirst. 2004. “The Aix-MARSEC Project: An Evolutive Database of Spoken British English.” In *Speech Prosody 2004*, International Conference; Nara, Japan, March 23–26, 2004, edited by Bernard Bel and Isabelle Marlien; ISCA Archive. http://www.isca-speech.org/archive/sp2004.

Beckman, M., and J. Pierrehumbert. 1986. “Intonational Structure in Japanese and English.” *Phonology Yearbook* 3:255–309.

Bigi, B., and D. Hirst. 2013. “What’s New in SPPAS 1.5?” In *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, edited by Brigitte Bigi and Daniel Hirst, 62–65. Aix-en-Provence, France.

Bigi, B., and D. J. Hirst. 2012. “Speech Phonetization Alignment and Syllabification (SPPAS): A Tool for the Automatic Analysis of Speech Prosody.” In *Speech Prosody 2012*, Sixth International Conference, Shanghai, China, May 22–25, 2012, edited by Qiuwu Ma, Hongwei Ding, and Daniel Hirst; ISCA Archive. http://www.isca-speech.org/archive/sp2012.

Bird, C. S., J. P. Hutchinson, and M. Kante. 1977. *Beginning Bambara (An Ka Ba-manankan Kalan)*. Bloomington: Indiana University Press.

Boersma, P., and D. Weenink. 1992 (2015). *Praat, a System for Doing Phonetics by Computer*. Version 5.4.06. Computer program. http://www.praat.org.

Bolinger, D. 1958. “A Theory of Pitch Accent in English.” *Word* 14:119–149.

Campione, E. 2001. “Etiquetage Semi-Automatique de la Prosodie dans les Corpus Oraux—algorithmes et Méthodologies.” PhD diss., Université de Provence.

Chentir, A., M. Guerti, and D. Hirst. 2009. “Extraction of Standard Arabic Micromelody.” *Journal of Computer Science* 5:86–89.

Cho, H., and S. Rauzy. 2008. “Phonetic Pitch Movements of Accentual Phrases in Korean Read Speech.” In *Speech Prosody 2008*, Fourth International Conference, Campinas, Brazil, May 6–9, 2008, edited by Plinio A. Barbosa, Sandra Madureira, and Cesar Reis; ISCA Archive. http://www.isca-speech.org/archive/sp2008.

Chomsky, N., and M. Halle. 1968. *The Sound Pattern of English*. New York: Harper.

Clements, G., and K. Ford. 1979. “Kikuyu Tone Shift and Its Synchronic Consequences.” *Linguistic Inquiry* 10:179–210.

Couper-Kuhlen, E. 1986. *An Introduction to English Prosody*. London: Edward Arnold.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

De Looze, C. 2010. “Analyse et Interprétation de L’empan Temporel des variations Prosodiques en Français et en Anglais.” PhD diss., Université de Provence Aix-en-Provence.

Di Cristo, A. 1998. “Intonation in French.” In *Intonation Systems*: *A Survey of Twenty Languages*, edited by D. J. Hirst, and A. Di Cristo, 195–218. Cambridge: Cambridge University Press.

Di Cristo, A., and D. J. Hirst. 1986. “Modelling French Micromelody: Analysis and Synthesis.” *Phonetica* 43 (1–3): 11–30.

Faure, G. 1962. *Recherches sur les Caractères et le Rôle des Éléments Musicaux dans la Prononciation Anglaise*. Paris: Didier.

Fujisaki, H. 1991. “Modeling the Generation Process of F0 Contours as Manifestation of Linguistic and Paralinguistic Information.” In *Proceedings of the Twelfth International Congress of Phonetic Sciences*. https://www.internationalphoneticassociation.org/icphs/icphs1991.

Fujisaki, H. 2004. “Information, Prosody and Modeling—with Emphasis on Tonal Features of Speech.” In *Speech Prosody 2004*, International Conference; Nara, Japan, March 23–26, 2004, edited by Bernard Bel and Isabelle Marlien, ISCA Archive. http://www.isca-speech.org/archive/sp2004.

Gårding, E. 1998. “Intonation in Swedish.” In *Intonation Systems*: *A Survey of Twenty Languages*, edited by D. Hirst and A. Di Cristo, 117–136. Cambridge: Cambridge University Press.

Goldsmith, J. A. 1990. *Autosegmental and Metrical Phonology*. Cambridge, MA: Blackwell.

Halliday, M. 1967. *Intonation and Grammar in British English*. Berlin: Mouton.

Hirst, D. J. 1974. "La Levée de L'ambiguïté Syntaxique par les Traits Intonatifs: Essai de Formalisation." PhD diss., Université de Provence.

Hirst, D. J. 1977. *Intonative Features*: *A Syntactic Approach to English Intonation*. Berlin: Mouton.

Hirst, D. J. 1983. "Interpreting Intonation: A Modular Approach." *Journal of Semantics* 2:171–181.

Hirst, D. J. 1987. "La Représentation Linguistique des Systèmes Prosodiques: Une Approche Cognitive." Thèse de Doctorat d'Etat (Habilitation Thesis), Université de Provence.

Hirst, D. J. 1998. "Intonation in British English." In *Intonation Systems*: *A Survey of Twenty Languages*. Edited by D. J. Hirst and A. Di Cristo, 56–77. Cambridge: Cambridge University Press.

Hirst, D. J. 2005. "Form and Function in the Representation of Speech Prosody." *Speech Communication* 46:334–347.

Hirst, D. J. 2007. "A Praat Plugin for Momel and INTSINT with Improved Algorithms for Modelling and Coding Intonation." In *Proceedings of the Sixteenth International Conference of Phonetic Sciences*, edited by Jürgen Trouvain and William J. Barry, 1233–1236. Saarbrücken, Germany.

Hirst, D. J. 2011. "The Analysis by Synthesis of Speech Melody: From Data to Models." *Journal of Speech Sciences* 1:55–83.

Hirst, D. J. 2012. "Empirical Models of Tone, Rhythm and Intonation for the Analysis of Speech Prosody." In *Rhythm*, *Melody and Harmony in Speech*: *Studies in Honour of Wiktor Jassem*, edited by D. Gibbon, D. J. Hirst, and N. Campbell, 23–33. Poznan, Poland: Polish Phonetic Association.

Hirst, D. J. 2015. "ProZed: A Speech Prosody Editor for Linguists, Using Analysis-by-Synthesis." In *Speech Prosody in Speech Synthesis*: *Modeling and Generation of Prosody for High Quality and Flexible Speech Synthesis*, edited by K. Hirose and J. Tao, 3–17. Berlin: Springer-Verlag.

Hirst, D. J. 2020. *Speech Prosody*: *From Acoustics to Interpretation*. Berlin: Springer-Verlag.

Hirst, D. J., and C. Auran. 2005. "Analysis by Synthesis of Speech Prosody: The Prozed Environment." In *Interspeech'2005—Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 4–8; 200 ISCA Archive. 3225–3228. Lisbon. http://www.isca-speech.org/archive/interspeech_2005.

Hirst, D. J., and C. Bouzon. 2005. "The Effect of Stress and Boundaries on Segmental Duration in a Corpus of Authentic Speech (British English)." In *Interspeech'2005—Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 4–8, 2005; ISCA Archive, 29–32. Lisbon. http://www.isca-speech.org/archive/interspeech_2005.

Hirst, D. J., H. Cho, S. Kim, and H. Yu. 2007. "Evaluating Two Versions of the Momel Pitch Modeling Algorithm on a Corpus of Read Speech in Korean." In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association,* Antwerp, Belgium, August 27–31, 2007; ISCA Archive, 1649–1652. Antwerp, Belgium. http://www.isca-speech.org/archive/interspeech_2007.

Hirst, D. J., and A. Di Cristo. 1984. "French Intonation: A Parametric Approach." *Die Neueren Sprachen* 83:554–569.

Hirst, D. J., and A. Di Cristo. 1998. *Intonation Systems*: *A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

Hirst, D. J., A. Di Cristo, and R. Espesser. 2000. "Levels of Representation and Levels of Analysis for the Description of Intonation Systems." In *Prosody*: *Theory and Experiment*: *Studies Presented to Gösta Bruce*, edited by M. Horne, 51–58. Amsterdam: Kluwer.

Hirst, D. J., and R. Espesser, R. 1993. "Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function." *Travaux de l'Institut de Phonétique d'Aix* 15:75–85.

Iivonen, A. 1998. "Intonation in Finnish." In *Intonation Systems*: *A Survey of Twenty Languages*, edited by D. Hirst and A. Di Cristo, 331–347. Cambridge: Cambridge University Press.

Jassem, W. 1952. *Intonation of Conversational English (Educated Southern British)*. Wroclaw, Poland: Wroclawskie Towarzystwo Naukow.

Jun, S.-A. 2005. *Prosodic Typology*: *The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Kingdon, R. 1958. *The Groundwork of English Intonation*. London: Longmans.

Ladd, D. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Liberman, M., and A. Prince. 1977. "On Stress and Linguistic Rhythm." *Linguistic Inquiry* 8:249–336.

Maghbouleh, A. 1998. "Tobi Accent Type Recognition." In *The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference*, Sydney Convention Centre, Sydney, Australia, 30th November–4th December 1998; ISCA Archive, Paper 0632. http://www.isca-speech.org/archive/icslp_1998.

Mixdorff, H.-J. 1999. "A Novel Approach to the Fully Automated Extraction of Fujisaki Model Parameters." In *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99*, Phoenix, Arizona, USA, March 15–19, 1999. IEEE Computer Society.

Nespor, M., and I. Vogel. 1986. *Prosodic Phonology*. Dordrecht, the Netherlands: Foris Publications.

Nooteboom, S. 1997. "The Prosody of Speech: Melody and Rhythm." In *The Handbook of Phonetic Sciences*, edited by W. Hardcastle and J. Laver, 640–673. Oxford: Blackwell.

O'Connor, J., and G. Arnold. 1961. *Intonation of Colloquial English: A Practical Handbook*. London: Longmans.

Palmer, H. E. 1924. *English Intonation with Systematic Exercises*. Cambridge: Heffer.

Pierrehumbert, J. B. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pike, K. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Rijsbergen, C. J. van. 1979. *Information Retrieval*. 2nd ed. London: Butterworth.

Selkirk, E. 1984. *Phonology and Syntax*: *The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. "TOBI: A Standard for Labeling English Prosody." In *Proceedings of the Second International Conference on Spoken Language Processing*, 867–870. Banff, Canada: ISCA.

Taylor, P. 1994. "The Rise/Fall/Connection Model of Intonation." *Speech Communication* 15:169–186.

Taylor, P. 2009. *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

’t Hart, J., R. Collier, and A. Cohen. 1990. *A Perceptual Study of Intonation*: *An Experimental-Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Trubetzkoy, N. S. 1949. *Grundzüge der Phonologie*. Translated by J. Cantineau (1957) as *Principes de phonologie*. Paris: Klincksieck.

Véronis, J., D. Hirst, and N. Ide. 1994. “NL and Speech in the MULTEXT Project.” In *Proceedings of AAAI Workshop on Integration of Natural Language and Speech*, 72–78. Seattle.

Wakefield, J. 2009. “The English Equivalents of Cantonese Sentence-Final Particles: A Contrastive Analysis.” PhD diss., Hong Kong Polytechnic University.

Wakefield, J. 2012. “A Floating Tone Discourse Morpheme: The English Equivalent of Cantonese lo1.” *Lingua* 122:1739–1762.

Wells, J. C. 2006. *English Intonation*: *An Introduction*. Cambridge: Cambridge University Press.

Wightman, C. 2002. “ToBI or Not ToBI”? In *Speech Prosody 2002, International Conference*; Aix-en-Provence, France, April 11–13, 2002; ISCA Archive, 21–29. Aix-en-Provence. http://www.isca-speech.org/archive/sp2002.

Wightman, C., and N. Campbell. 1995. “Improved Labeling of Prosodic Structure. In *Proceedings of IEEE Transactions on Speech and Audio Processing*. Piscataway, NJ: IEEE.

Xu, Y., and X. Sun. 2002. “Maximum Speed of Pitch Change and How It May Relate to Speech.” *Journal of the Acoustical Society of America* 111:1399–1413.

Xu, Y., and Q. E. Wang. 2001. “Pitch Targets and Their Realization: Evidence from Mandarin Chinese.” *Speech Communication*, 33:165–181.

Zhi, N., D. Hirst, and P. M. Bertinetto. 2010. “Automatic Analysis of the Intonation of a Tone Language. Applying the Momel Algorithm to Spontaneous Standard Chinese (Beijing).” In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, Makuhari, Japan, September 26–30, 2010, edited by Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura; ISCA Archive. http://www.isca-speech.org/archive/interspeech_2010.

4

The ToBI Transcription System: Conventions, Strengths, and Challenges

Sun-Ah Jun

4.1 Introduction

Decisions regarding what information a prosodic transcription should include vary depending on the goal of transcription. For example, the goal may be to (i) represent the distinctive features of the prosodic system of a language, (ii) learn how prosodic categories are phonetically realized and how variable the categories are, (iii) train machines to achieve better synthesis and recognition of human speech, (iv) compare prosody across dialects or across languages, or (v) develop a corpus to explore various linguistic phenomena related to prosody. Thus, prosodic transcriptions may include annotations of prosodic properties that are meaningful at the level of phonology, phonetics, or even lower (i.e., the implementation of phonetic rules for a specific language), or they can be nondistinctive but still categorical in nature, capturing prosodic properties not specific to a certain language.

ToBI, which stands for tones and break indices, is the most well-known annotation system for prosody at the level of phonology. It aims to transcribe the phonological properties of intonation (the "tones" part) and the perceived degree of juncture between words (the "break indices" part), which together represent the prominence patterns and prosodic structure of an utterance. Thus, the tone labels used in the ToBI system of a specific language are meant to represent underlying tonal targets that are both meaningful to native speakers of the language and systematic and consistent across native speakers of the language variety.

The ToBI system was originally designed for transcribing the intonation and prosodic structure of English utterances (Silverman et al. 1992; Beckman and Hirschberg 1994; Beckman and Ayers Elam 1997; Beckman, Hirschberg, and Shattuck-Hufnagel 2005), and so it began as a transcription system for the prosody of English, more specifically for mainstream American English. However, after the development of the ToBI systems for other dialects of English, such as GlaToBI for Glasgow English (Mayo, Aylett, and Ladd 1997), as well as a few typologically different languages—for example, GToBI for German (Grice and Benzmüller 1995), K-ToBI for Korean (Beckman and Jun 1996; Jun 2000), and J_ToBI for Japanese (Venditti 1997), *ToBI* has come to refer to a general framework for prosodic transcription systems based on phonological properties.

The original ToBI was therefore renamed MAE_ToBI (Mainstream American English ToBI, although sometimes it is still simply called "English ToBI") when several ToBI systems of typologically various languages (e.g., Bininj Gun-wok, Cantonese, Chickasaw, German, Greek, Japanese, Korean, Mandarin, Serbo-Croatian) were presented at a satellite meeting of the Fourteenth International Congress of Phonetic Sciences in San Francisco in 1999.[1] Since then, many ToBI systems have been developed, covering languages whose prosodic systems had been rarely studied or studied in different frameworks, such as Spanish (Beckman et al. 2002; Face and Prieto 2007), Bangladesh Bengali (Khan 2008, 2014), Mongolian (Indjieva 2009), Catalan (Prieto et al. 2009; Prieto 2014), European Portuguese (Frota 2014; Frota et al. 2015), and French (Delais-Roussarie et al. 2015), among others.

This chapter introduces what the ToBI conventions are and how the ToBI transcription system works and discusses the strengths and weaknesses of the ToBI system. It also describes some of the recent developments in the prosodic transcription system that attempt to address the ToBI system's known limitations. Before introducing how the ToBI system works, section 4.2 offers a brief description of the theoretical background and framework on which ToBI is based. Section 4.3 presents the ToBI conventions and how the ToBI system has been applied to typologically various languages, showing the workings of the phonological theory it has adopted. Section 4.4 presents the strengths of the ToBI system; section 4.5 discusses the problems and challenges ToBI users face, as well as recent developments that have been made in response to such challenges; and section 4.6 concludes the chapter.

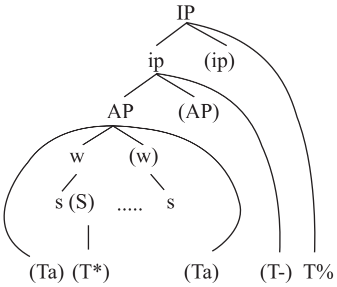
4.2 The Theoretical Background of the ToBI System

The ToBI system includes two main types of annotation: tonal labels and break index labels. The tonal labels are based on the autosegmental-metrical (AM) model of intonational phonology. (See Bruce 1977; Liberman 1975; Pierrehumbert 1980; Liberman and Pierrehumbert 1984; Beckman and Pierrehumbert 1986; Pierrehumbert and Beckman 1988; Ladd 1983, 2008a; and Pierrehumbert, commentary to chapter 11, this volume. See also Arvanti, chapter 1, this volume, for a detailed description of the AM model of intonational phonology.) In this model, an intonational contour is analyzed as a linear sequence of high (H) and low (L) tones. As an autosegment, the intonational tones can be associated with a specific syllable or mora in a word, that is, head, with the edge of a specific prosodic unit, or both, reflecting the metrical and/or prosodic structure of the utterance (Beckman 1996). When the tone is associated with an intonationally prominent syllable (either because it is metrically or rhythmically strong or because it is lexically marked), it is called a *pitch accent* (marked with a star [*], e.g., H*, following Goldsmith 1976 and Pierrehumbert 1980). When the tone is associated with the edge of a prosodic unit, it is called a *boundary tone* (e.g., H%, following Liberman 1975 and Pierrehumbert 1980). In this way, these intonational tones achieve two major functions: (i) marking the prominence relationship among the syllables (or moras) within a word and among the words within a phrase and (ii) marking prosodic grouping and hierarchical prosodic structure of the utterance.[2]

The AM model of intonational phonology, illustrated well in the models proposed by Beckman and Pierrehumbert (1986) and Pierrehumbert and Beckman (1988), assumes three prosodic units defined by intonation: an *intonational phrase* (IP), which is the largest prosodic unit; an *intermediate phrase* (ip), a prosodic unit smaller than an IP; and an *accentual phrase* (AP), which is slightly larger than a word (w) and smaller than an ip. These prosodic units form a strictly layered hierarchical prosodic structure (Selkirk 1986; Nespor and Vogel 1986; Pierrehumbert and Beckman 1988). A higher-level prosodic unit is assumed to be exhaustively parsed into one or more lower-level prosodic units. Therefore, any word or syllable should be part of each prosodic unit, and because an IP dominates an ip, which dominates an AP, an IP-final syllable is assumed to also be an ip-final syllable, as well as an AP-final syllable. A tree diagram of intonationally defined prosodic structure is shown in (1).[3] Each prosodic unit can be marked by a boundary tone (T% for an IP boundary tone, T- for an ip boundary tone, Ta for an

AP boundary tone; where T = H or L or its combinations), which is typically realized on the last syllable of a prosodic unit, regardless of whether the syllable is stressed. Among the boundary tones, only an IP boundary tone is obligatory. Lowercase *s* refers to a syllable, and *S* is a metrically strong/lexically marked syllable (the latter of which can be associated with a pitch accent, T*).

(1) Intonationally defined prosodic structure.



The AM model of intonational phonology allows only two tonal levels, H and L, and as in the analysis of African tone languages (Goldsmith 1976; Leben 1976), a rising or falling contour tone is represented by a combination of these level tones. This differs from the British intonation models (e.g., Palmer 1922; Halliday 1967; Crystal 1969) and from Bolinger's (1951, 1958) model where a contour tone (configuration) was represented as a single tonal unit. That is, a pitch accent in the AM model can be monotonal or bitonal, depending on the shape of the fundamental frequency (F0) contour around an accented syllable. A monotonal pitch accent represents a level tone, such as H* and L*, and a bitonal pitch accent represents a rising tone such as L + H* or a falling tone such as H* + L. By decomposing a contour tone into an L and an H level tone, the AM model can also distinguish contour tones where the F0 peak or valley is aligned differently relative to an accented syllable. That is, when a bitonal pitch accent is associated with an accented syllable of a prominent word, the starred tone of the pitch accent is realized on the accented syllable of the word. The unstarred tone, which either precedes (called a *leading tone*) or follows (called a *trailing tone*) the starred tone, is typically realized on the syllable immediately preceding or following the star-toned syllable, respectively. For example, L + H* is a rising tone with an F0 peak during the accented syllable and an F0 valley on the immediately preceding syllable, while L* + H is a rising tone with an F0 valley during the accented syllable and an F0 peak on the immediately following syllable.

As a phonological model of intonation adapted mostly from Pierrehumbert (1980), the AM model of intonational phonology posits the simplest possible abstract underlying representation. It accounts for the surface variation of underlying tones by rules that map the phonological representation (abstract level tone target sequences) to the phonetic representation of intonation (the F0 contour). Thus, even though the AM model assumes only two tonal levels in the underlying representation, various tonal levels realized in the surface F0 contour are explained by the rules that affect pitch range, such as upstep or downstep relative to the immediately preceding tone, the strength of prominence (e.g., a high nuclear pitch accent is often higher than a high prenuclear pitch accent in English), the lexical status of a tone (e.g., a lexical H tone is higher than a phrasal H tone in Japanese), or the function of tone (e.g., an L% boundary tone is typically lower than a L* pitch accent).

In addition to the rules adjusting the pitch range for a tonal category's realization, the model also adopts the target-interpolation rule to explain the surface F0 contour over syllables that have no underlying tonal targets. As a phonological model, only distinctive pitch targets are associated with a certain syllable, and syllables without a tonal target receive their surface F0 values through direct interpolation between two adjacent tonal targets, that is, the tones before and after the toneless syllable(s). In general, interpolation is done locally between two tonal targets, but it can be sensitive to the tone's function. For example, in English, interpolation does not occur between a prominence-marking pitch accent and a boundary tone in the same way that interpolation occurs between two pitch accents.

Distinctive intonational tone categories, such as pitch accents and boundary tones, proposed for a specific variety of a language will form the tonal inventory for that variety's ToBI system. But a complete ToBI annotation system should also include labels for the break index (BI) of the language. The BI in the ToBI system is in general marked by Arabic numerals, representing the perceived degree of juncture between any two adjacent words. Typically, the higher the numeral, the larger the perceived juncture between the words. Therefore, the distribution of BIs reflects prosodic grouping of words and thus the utterance's hierarchical prosodic structure.

The convention of using numbers to represent the degree of juncture is adopted from work by Price and her colleagues (Price et al. 1991; Wightman et al. 1992), who examined the mapping between syntactic and prosodic structure, and the role of prosody (especially boundary marking and prominence) in the resolution of various types of syntactic ambiguity. As Beckman et al. (2005) described in detail, the ideas that Price and her colleagues investigated have a long tradition, going back to the work known as *instrumental phonetics* (see Ladd [1996] 2008a and Price et al. 1991 for a review), where researchers examined the phonetic correlates of syntactic structure (e.g., Lieberman 1967; Lehiste 1973; Klatt 1975; Cooper and Paccia-Cooper 1980; Scott 1982), and work on *prosodic phonology* (e.g., Gee and Grosjean 1983; Selkirk 1986; Nespor and Vogel 1986), where researchers tried to predict prosodic structure from syntactic structure and explained the domain of postlexical phonological rules based on a sentence's syntactic structure.

In sum, the ToBI system is a transcription system for annotating the phonological properties of prosody, especially the prominence and prosodic structure of an utterance, analyzed in the AM model of intonational phonology for a specific language variety. The following section describes the conventions of the ToBI system based on data from English and other typologically various languages, illustrating how the phonological approach of AM theory is manifested in various ToBI systems.

4.3 ToBI Conventions and Various ToBI Systems

The goal of developing the original ToBI system was to provide a common vocabulary for tagging phonological properties of prosody and intonation to a community of users working at different sites. This common vocabulary would then facilitate the ability to share and interpret each other's data and to build a large prosodically transcribed database. In their chapter describing the original ToBI system, Beckman et al. (2005, 12–14) state that a viable ToBI system must conform to the following four principles. First, the conventions should be "accurate," conforming to an established body of research in the intonational phonology of the language variety, and, when possible, the conventions should also be informed by work on dialectology, pragmatics, and discourse analysis for the language. Second, the conventions should be "efficient," by not transcribing information that can be extracted from the signal or derived from online resources automatically (e.g., location of lexical stress). Also, prosodic phenomena should not be labeled if they cannot be labeled consistently due to a lack of understanding of the phenomena by the research community. Third, the conventions should be "easy" enough to learn by consulting a manual (which includes sound files illustrating ToBI transcriptions), and "reliable" enough to be applied consistently among transcribers. For this, intertranscriber consistency should be evaluated and the conventions should be reviewed regularly and upgraded, if needed, by agreed-on groups. Finally, a ToBI transcription should never replace a permanent record of the speech signal with a symbolic record. Instead, it should integrate symbolic labels with continuous measures derived directly from the signal. This is because a ToBI system is not simply a transcription system, but also "a tool for observing the signal and communicating one's observations to the larger community in a common language" (Beckman et al. 2005, 14).

A ToBI transcription therefore requires an audio recording of an utterance and a record of the F0 contour, aligned with symbolic labels written on four parallel quasi-independent tiers: words, tones, BIs, and miscellaneous (misc). In the original ToBI system, UNIX-based xwaves software was used to display the waveform, pitch track, and four tiers (placed above the pitch track), and a label on each tier was placed right-justified relative to a certain time point. That is, the end of each label is aligned with the end of the acoustic signal corresponding to each word (for the words and BIs tiers) or with the time point of a tonal event (for the tones tier) or a nonspeech event or interval (for the misc tier). Figure 4.1 provides an example, showing the English utterance "Okay, they have a couple flights," transcribed in MAE_ToBI, using xwaves. Later ToBI systems have used different software, such as PitchWorks (Scicon R&D; http://www.sciconrd.com/index.aspx), WaveSurfer (Department of Speech, Music, and Hearing, KTH; http://www.speech.kth.se/wavesurfer/), or Praat (http://www.fon.hum.uva.nl/praat/). Praat has been used widely in recent years by ToBI users, and unlike the format used in xwaves, has two different types of labeling tiers: it places a label either as an event that has some duration, in an *interval tier*, or as an event occurring at a specific time point, in a *point tier*. In this format, an interval tier is appropriate for the words tier, which marks the beginning and the end of each word determined by examining the waveform and spectrogram, and a point tier is appropriate for the other three tiers. Figure 4.2 shows an example of ToBI transcription of the same utterance as in Figure 4.1, but displayed in Praat. The subsections that follow describe the conventions on each tier based on data from English and other languages.

4.3.1 The Words Tier

The words tier is where an orthographic transcription of each word in the utterance appears, aligned with the acoustic representation of the word (in the waveform or the spectrogram, or both). What counts as a word is language-specific, and how to transcribe it can be decided by the research community for a specific ToBI system. For example, even though the morphosyntactic function of a case marker or a postposition is similar between Japanese and Korean, these morphemes are treated differently prosodically in each language. They are entered as an independent word in J_ToBI (aligned with BI 1), but as part of a word in K-ToBI. This is because case markers and postpositions are sometimes produced separately from a stem in Japanese, but not in Korean.

The orthography of a word can be written in language-specific characters if the font is available, or in a romanized text of the word, as in J_ToBI or K-ToBI. In Pan-Mandarin

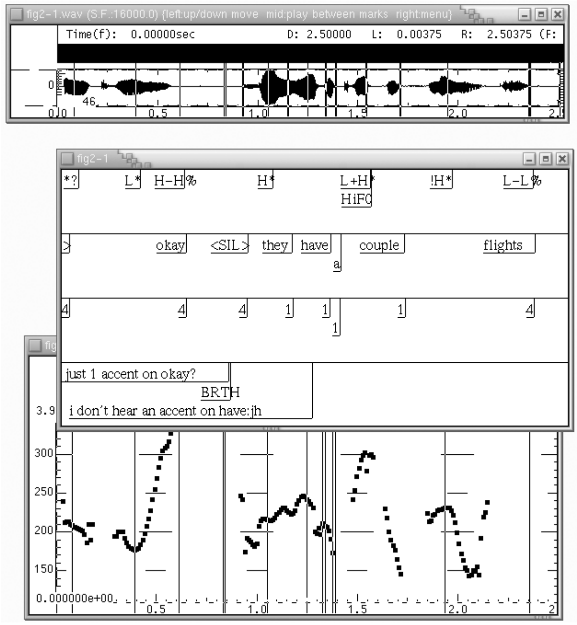


fig2-1.wav (S.F.:16000.0) {left:up/down move mid:play between marks right:menu}
Time(f): 0.00000sec D: 2.50000 L: 0.00375 R: 2.50375 (F:
0
46
0.0
0.5
1.0
1.5
2.0
fig2-1
*?
L*
H-H%
H*
L+H*
HiF0
!H*
L-L%
>
okay
<SIL>
they
have
a
couple
flights
4
4
4
1
1
1
1
4
BRTH
3.9
300
250
200
150
0.000000e+00
0.5
1.0
1.5
2.0
2

just 1 accent on okay?

i don't hear an accent on have:jh

Figure 4.1
An example English utterance, "Okay. They have a couple flights," transcribed in MAE_ToBI in xwaves. The tiers are shown in the order of words, tones, BIs, and miscellaneous. *Source*: Courtesy of Beckman, Hirschberg, and Shattuck-Hufnagel (2005, figure 2.1); reproduced by permission of Oxford University Press. https://global.oup.com/academic/product/prosodic-typology-9780199208746?cc=us&lang=en&. BRTH, breath.

ToBI (Peng et al. 2005), syllable-by-syllable transcriptions in Chinese characters are given on the words tier, but transcription of syllables in modified Pinyin romanization is given on the romanization tier, called the *Romazi tier*. Similarly, the words tier in Cantonese C-ToBI (Wong, Chan, and Beckman 2005) provides a syllable-by-syllable transcription in Chinese characters, but because not every labeling platform has Chinese characters, the C-ToBI developers suggest making the words tier optional and a syllables tier obligatory, where an alphabetic transliteration is provided for every syllable in the words tier.

Finally, when the boundary of a word is not easy to locate (for example, due to cliticization of adjacent function words), the orthography of a prosodic word rather than a morphosyntactic word (e.g., *gonna* rather than *going to*) can be written on the words tier.

4.3.2 The Tones Tier

The tones tier is for transcription of distinctive tonal events and other tone-related tags. Distinctive tonal events can include (i) phrasal tones, marking the edge of a prosodic unit (e.g., Ha for an accentual phrase boundary tone in Bengali [Khan 2008, 2014] and Korean [Jun 2000, 2005a] or L% for a low boundary tone of an IP in many languages); (ii) postlexical pitch accents, marking the prominence of a word (e.g., L*, L+H* in English, German [Grice, Baumann, and Benzmueller 2005], Greek [Arvaniti and Baltazani 2005], or Bininj Gun-wok [Bishop and Fletcher 2005]); or (iii) lexically specified tonal events (e.g., lexical pitch accent H*+L in Japanese [Venditti 1997, 2005] and

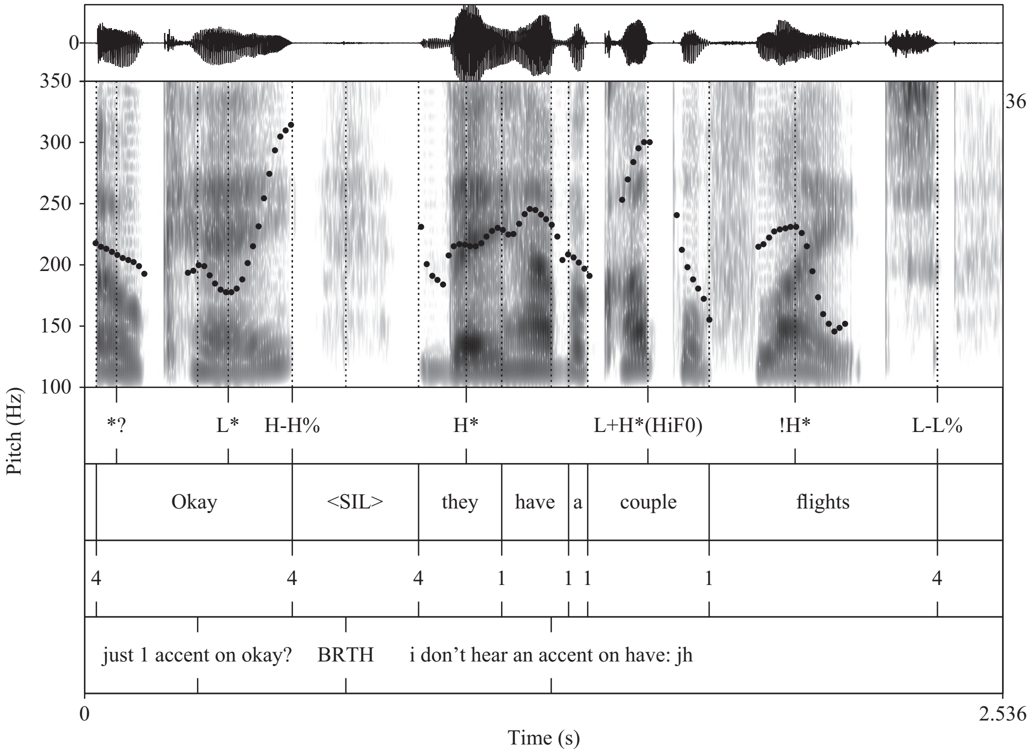


Figure 4.2
An example of MAE_ToBI transcription of the same utterance shown in figure 4.1 but transcribed in Praat.

L* + H in Serbo-Croatian [Godjevac 2005]), lexical tones in Cantonese [Wong, Chan, and Beckman 2005], or H^{λ} morphological pitch accent in Chickasaw [Gordon 2005]).

Because distinctive tonal categories and their realizations are language- or dialect-specific, the same tonal label is not necessarily realized in the same way across languages, and the same F0 contour can be transcribed with different tonal labels across languages. For example, as mentioned earlier for English intonation, a rising pitch accent L* + H is used to label an F0 contour with an F0 valley on an accented syllable and an F0 peak on the immediately following syllable. However, in Greek, the same tonal label is used for an F0 contour with an F0 valley at the very beginning or slightly before the onset of the accented syllable and an F0 peak at the beginning of the postaccentual vowel (Arvaniti, Ladd, and Mennen 1998, 2000; Arvaniti and Baltazani 2005). This is possible because each of these languages has another, contrasting rising pitch accent, such as L + H*, whose F0 peak is aligned later than that of L* + H. Thus, what justifies a tonal label in a language is its status as contrastive or noncontrastive, not the exact details of its phonetic realization.

Furthermore, the same F0 contour with the same tone-text alignment can be transcribed differently across languages depending on the lexical prosody of the target language. For example, a phrase-medial rising F0 contour over a two syllable word can be categorized as either L* + H or L + H* if the first or the second syllable, respectively, is the metrical head of the word (as in English, German, and Spanish ToBIs). If both tones are not associated with a head syllable, the contour can be categorized as a LH boundary tone or a LH tonal melody (as in Mongolian and Korean ToBIs) or an L boundary tone followed by an H phrasal one (as in J_ToBI). Or, if the first syllable is the head syllable of a word while the following syllable is the last syllable of a word, the sequence can be categorized as an L* pitch accent followed by an H boundary tone (as in Bangladesh Bengali ToBI).

In addition to the difference in tone-text alignment, tone labels reflect differences in the tonal scale or pitch range. While the AM model assumes only two tonal levels, H and L, it is also known that an H tone target is sometimes lower compared to the preceding H target, more so than would be expected given simple declination over the utterance. Furthermore, such reduction in pitch range may apply iteratively to the subsequent H tone targets. H tones realized in such a reduced pitch range are called "downstepped H tones." Because the context of downstep is similar to the downdrift phenomenon in African tone languages (Leben 1973), Beckman and Pierrehumbert (1986) proposed downstep for their AM model of English intonation to be triggered by the preceding bitonal pitch accent within the same intermediate phrase. However, the original ToBI conventions for English, influenced by Ladd (1983), adopted a theory-neutral downstep diacritic, !, before H to annotate a downstepped H tone. With this change, the MAE_ToBI revised the tonal label H+L* in Beckman and Pierrehumbert (1986) to H+!H* to reflect an F0 fall from a high to mid target rather than from a high to low target, and revised the label H*+L to H* followed by !H*. This therefore increased the tone inventory for English by adding a downstepped version of all pitch accents that have an H tone component (i.e., !H*, L+!H*, L*+!H, !H+!H*) and also a downstepped version of the H phrase accent (i.e., !H-).[4] This introduction of a downstep diacritic made the tonal category in the ToBI system more faithful to the surface F0 contour than that proposed in the AM model.[5]

The tones tier also includes tones marking the prosodic structure of an utterance. As mentioned in section 4.1, a language can have at least three prosodic units marked by intonation above the word level: IP, ip, and AP.[6] The tones marking the edge of a prosodic unit, called *boundary tones*, are accompanied by a diacritic, indicating that the prosodic unit that the tone is affiliated with. For example, *a* is used to mark an AP boundary tone (e.g., Ha, La), - for an ip boundary tone (e.g., H-, L-), and % for an IP boundary tone (e.g., H%, L%). Unlike the pitch accent, which is located within the accented syllable, specifically aligned with the F0 minimum for L*-type tones (e.g., L*, L*+H) and with the F0 maximum for H*-type tones (e.g., H*, L+H*), the boundary tone is typically located on the tones tier, aligned with the end of the relevant prosodic unit. When an ip is IP-final, though, both ip-boundary and IP-boundary tones can be realized sequentially at the edge of an IP, as in English (e.g., L-H%). In some languages, however, the boundary tone marking a lower-level unit (e.g., an ip) can be overridden by that marking a higher-level unit (e.g., an IP), as in Bangladesh Bengali or Korean. Thus, how multiple boundary tones are realized when they are hosted by the same syllable is language-specific.

The pitch accents and boundary tones mentioned so far are all postlexical tones, but the tones forming an intonation contour can also originate from a lexical tone, such as a lexical pitch accent (in Japanese and Swedish), a morphological pitch accent (in Chickasaw), or a lexical tone (in Mandarin and Cantonese). In the ToBI system of most languages, both lexical and phrasal tones are labeled on the same tier, a tones tier, because the lexical status of the tone is recognizable from the diacritics on the tone symbol. For example, in J_ToBI (1997, 2005), there is only one lexical pitch accent and it is marked with a star diacritic (i.e., H* + L). All other tones are phrasal tones: an AP phrasal tone marked with a diacritic - (i.e., H-) or a boundary tone marked with a diacritic % (i.e., an L% AP-boundary tone and a few IP-boundary tones [e.g., H%, LH%]). In Chickasaw ToBI (Gordon 2005), there is one morphological pitch accent, and it is marked with a lambda (i.e., H^{λ}). All other tones are phrasal tones: a postlexical pitch accent marked with a star (e.g., H*), AP phrasal tones without any diacritics (e.g., L, H), and an IP boundary tone marked with % (e.g., H%). Similarly, both lexical tones and IP-boundary tones are labeled on the tones tier in Cantonese ToBI (Wong, Chan, and Beckman 2005), but lexical tones are represented using Chao's system of tone numbers (e.g., 35, 53) and IP tones are marked with % (e.g., L%, HL%). However, lexical tones and postlexical tones can also be labeled on two separate tiers. In the Pan-Mandarin ToBI (Peng et al. 2005), the lexical tone is labeled on the Romazi tier while postlexical tonal events such as IP-boundary tones (i.e., L%, H%) and pitch-range–related labels (e.g., %reset, %e-prom) are labeled on the tones tier.

Finally, as shown in the original ToBI system, the tones tier can also include tone-related tags that are not intended to mark distinctive tonal events. Such tags are instead used to mark surface F0 patterns related to the phonetic realization of a tonal category or to discourse information. The most common example is to use a diacritic to mark atypical or allotonic realizations of an underlying tonal category, such as the use of < to mark a delayed F0 peak and > for an early F0 peak when the F0 peak is realized later or earlier than expected, respectively, from the typical tone-text alignment of the underlying tone. Another example is the use of *w* to mark a weak realization of an underlying tonal target, as in the undershoot of an L target in Greek prenuclear pitch accents (i.e., wL* + H) or in Japanese AP boundary tones (wL% or %wL).[7] Tone-related tags can also include a label that can be used to retrieve F0 values to study discourse structure (e.g., HiF0 in English on the highest F0 point among pitch accents within an ip), or a label to capture how pitch range is influenced by disfluency (e.g., %r at the onset of pitch-range reset after disfluency). Because a ToBI system is a tool for observing the signal and creating a communal corpus, adding these tags on surface F0 information related to the distinctive tonal categories allows researchers to learn how an underlying tone is phonetically realized in what context and how intonation interacts with other components of grammar and reflects various communicative pressures.

This idea is further extended in K-ToBI, version 3 (Jun 2000, 2005a), where the tones tier is split into two tiers, a phonological tone tier and a phonetic tone tier, to capture the variability and distribution of the surface tonal categories of the AP. Unlike English intonation, where all tonal categories are distinctive and each type of pitch accent contributes compositionally to the pragmatic interpretation of an utterance (Pierrehumbert and Hirschberg 1990), most of the tonal targets in Korean intonation are not distinctive. Distinctive tones are those marking the end of prosodic units larger than a word, and tones that are not AP-final are not distinctive. For example, an AP-initial tone can be L or H depending on the laryngeal feature of the AP-initial segment,[8] that is, predictable and thus not distinctive. AP-internal tones can be H on the second syllable or L on the penultimate syllable, but the presence of these tones varies by factors not fully known, though some of the tendencies can be explained by the AP length and by speech style. Even the AP-final tone, though usually H, can be L for factors that are also not well understood. Because the tones in ToBI systems are intended to be distinctive, nondistinctive yet categorical F0 targets such as these should not be included in the tones tier. For this reason, Jun (2000) chose to place these on a separate tier called the phonetic tone tier and put only distinctive tones on the phonological tone tier for

K-ToBI. It was hoped that labeling these surface tonal categories in a large corpus will help researchers to determine what factors and conditions trigger such variation, and thus improve the current model of Korean intonational phonology.

In sum, the tones tier is used to transcribe distinctive tonal categories of a specific language variety, regardless of their function (marking prominence or prosodic structure) and lexical status (lexical or phrasal), as well as to transcribe surface tonal events that are meaningful and interesting to the research community of the target language.

4.3.3 The Break Indices Tier

The BIs tier is for transcribing the perceived degree of juncture or the strength of association between each pair of words (and between the final word and the silence at the end of the utterance), capturing the hierarchical nature of an utterance's prosodic grouping. The original ToBI system (for English) proposed five indices in numerical numbers: four numbers for four different degrees of disjuncture and one number for a mismatch between the degree of juncture and the tonal cues that mark prosodic grouping. Among the four degrees of juncture, two corresponded to prosodic boundaries higher than the word level, that is, IP and ip, and two lower than the ip-level juncture, that is, ip-medial word boundary and a boundary weaker than a typical word boundary. (2) shows a brief description of five BIs in MAE_ToBI, represented in numeric numbers from 0 to 4. On the BIs tier, each index is labeled and aligned with the end of each word in the words tier.

(2) The BIs in MAE_ToBI

0 A juncture smaller than a typical phrase-medial word boundary. It corresponds to the juncture between a content word and a function word when the function word is cliticized. It can also be used for a weakened word juncture due to coalescence of adjacent segments across a word boundary.
1 A typical phrase-medial word boundary.
3 A juncture corresponding to an ip boundary.
4 A juncture corresponding to an IP boundary.
2 A mismatch between a tonal cue and the degree of juncture that mark a prosodic boundary. It includes two types of mismatch: (i) a juncture corresponding to ip or IP but with no tonal event defining ip or IP, and (ii) a juncture weaker than an ip level but with a clear tonal event for an ip or IP.

Because the BI is tightly linked to the prosodic structure of a language, the number of BI categories will differ across languages, reflecting the number of prosodic units a language has. However, the number of prosodic units defined by a tone does not necessarily match the number of BIs, because a prosodic unit is not necessarily marked by intonation. For example, Cantonese has only one prosodic unit defined by intonation, IP, but its ToBI system has three BI numbers: 2 for an IP boundary, 1 for a foot boundary, and 0 for a foot-internal syllable boundary. Furthermore, the Pan-Mandarin ToBI has six BI numbers, but only one BI (BI 4) corresponds to a prosodic unit defined by pitch reset.

As the list of BIs shows, an increasing value of BI refers to an increasing degree of juncture: $0 < 1 < 3 < 4$. For this reason, BI 2, which refers to a mismatch in the original ToBI, has sometimes been misinterpreted as a juncture larger than BI 1 and smaller than BI 3. Furthermore, the BI 2 in MAE_ToBI included two types of mismatch cases, as described in the BIs list in (2). To improve this situation, some of the later ToBI systems (e.g., Japanese, Korean, Greek, and Chickasaw) used BI 2 as the juncture larger than BI 1 and smaller than BI 3, and added a diacritic *m* after a BI number for mismatch cases. For example, 2m is used to label a BI 2–like juncture but without having tonal cues for the prosodic unit corresponding to BI 2 (which is an AP in Japanese, Korean, and Chickasaw, but an ip in Greek). Similarly, 3m is used to label a BI 3–like juncture but without having tonal cues for the prosodic unit corresponding to BI 3 (which is an IP in all four languages mentioned).

Because some of the BIs do correspond to a tonally defined prosodic unit, labeling both a BI on the BIs tier and a boundary tone on the tones tier would mark the same prosodic unit. For example, in MAE_ToBI, BI 3 and BI 4 are labeled on the BIs tier aligned with a phrase accent and an IP boundary tone, respectively, on the tones tier. This seems to violate the "efficiency" principle of the ToBI system, mentioned at the beginning of section 4.3. Namely, the system has to be efficient and should not transcribe predictable information. So, the researchers of a specific ToBI system could decide not to label BIs that are predictable from other tiers but label only those not predictable such as mismatch cases. But this would work if both tones and BI tiers are fully specified. In the cases where a research group wants to examine only the tonal patterns of a prosodic structure or only the degree of juncture or grouping between words, the tier of interest should be fully labeled regardless of whether some label is predictable from the other tier. That is, depending on the goal of the researchers, what to label on what tier can be different.[9]

Finally, a BI can also transcribe a disfluent juncture, a juncture produced disfluently or unnaturally for various reasons (e.g., a memory problem, a sudden change in a word choice, interruption by other speakers, stuttering), and the original ToBI proposed three types of labels for a disfluent juncture, as in (3).

(3) ToBI labels for disfluent juncture
1p for an abrupt cutoff of a word
2p for prolongation, but without a phrase accent
3p for prolongation, with an accompanying phrase accent

These two types of disfluency, cutoff and prolongation, have been adopted by most of the later ToBI systems, but these are not enough to transcribe all types of disfluencies found in speech.[10] To transcribe stuttered speech, Arbisi-Kelm (2006) increased the types of disfluency from two to five by adding "restart," "pause," and "filler." However, BIs describing disfluent junctures have not been used extensively across ToBI systems, probably because most ToBI systems have been proposed based on reading and scripted speech or semi-spontaneous speech and not much from free conversation speech.

4.3.4 The Miscellaneous Tier

The misc tier is for transcription of any comments or remarks about the utterance. It can be a comment about the utterance such as silence, audible breaths (an example is "BRTH" in figure 4.1), laughter, false start, hesitation, disfluencies, and other spontaneous speech effects. It can also be used to include labelers' remarks for themselves about a transcription, or questions to other labelers (e.g., "just 1 accent on okay?" in figure 4.1). If a label on the misc tier is for a localized event, it should be located at the time point of the event (e.g., a label "disfl" is written at the approximate time point where some disfluency is perceived), but if the event has an identifiable interval such as laughter does, the convention is to use paired labels to mark the starting point of the event with < and the ending point of the event with > (e.g., laughter< . . . laughter>).

4.3.5 ToBI Labels for Uncertainty and an Alt Tier

The original ToBI system provided a convention to transcribe a labeler's uncertainty about a BI or a tonal category. When a labeler is not sure about the degree of juncture between two adjacent BIs, the convention is to add a dash (-) after a BI of a larger juncture (e.g., 4- is used for a BI ambiguous between 3 and 4).[11] For a tonal category, *? is used when a labeler is not sure whether a syllable is pitch accented or not, and X*? is used when a labeler is sure about the presence of pitch accent on a syllable but not sure about the pitch accent type. The same rule is applied to other tonal categories. So, uncertainty about the presence and the type of a phrase accent is labeled as -?and X-?, respectively, and that of an IP boundary tone is labeled as %?and X%?, respectively.

However, this way of labeling uncertainty is often found to be unsatisfactory because it does not record relevant details of the ambiguity. That is, while it tags that the labeler is uncertain about the transcription, it does not specify what possible tonal labels were considered by the labeler, or whether one potential label should be preferred over another. To fix this problem, a fifth tier, Alt (alternative), was proposed at the Workshop on ToBI for Spontaneous Speech, held in Boston in 2004, and the use of an Alt tier has been tested by labeling a large speech sample in Brugos et al. (2008). Brugos and her colleagues proposed adding a question mark next to an ambiguous or uncertain tonal label (e.g., H*?) or a BI label (e.g., 3?) in the respective tier and providing an alternative tonal or BI label on the Alt tier (e.g., !H* or 2), with the alternative label aligned with the corresponding labels on the tones or BIs tiers. If a sequence of labels is ambiguous, instead of one label, they proposed adding a square bracket before and after the affected labels on the Alt tier (e.g., [2 L+H*]).

4.3.6 Other Tiers

As mentioned, the original ToBI system provided four core tiers (tones, words, BIs, misc), but later ToBI systems have added new tiers either to capture language-specific prosodic properties (e.g., Romazi, syllable, stress, and sandhi tiers in Pan_Mandarin ToBI; foot and syllable tiers in C_ToBI), to study the interface between prosody and other areas of linguistics (e.g., the phonetic transcription tier in GRToBI and Chickasaw ToBI, the prosodic word tier in GRToBI, the finality tier in J_ToBI), or to help labelers transcribe non-native or unfamiliar languages (e.g., the gloss tier in Chickasaw ToBI and Bininj Gun-wok ToBI; see Jun 2005b for examples of each tier used in different ToBI systems). Creation of these noncore tiers in various languages exemplifies the flexibility and extendibility of ToBI system as a tool for transcribing language-specific prosodic features as well as other linguistic properties that are useful to the research community or system users.

4.4 Strengths of the ToBI System

The strengths of the ToBI system come from the transcription of tones being phonological, especially based on the AM model of intonational phonology, and from the architecture and mechanism of the transcription system.

First, as a phonological transcription system based on the AM framework of intonational phonology, the ToBI system transcribes distinctive tonal categories of intonation that mark prominence and a hierarchical prosodic structure in a specific language variety. That is, the tones that are transcribed are not just a sequence of F0 turning points on the surface F0 contour, but are contrastive in the language by performing linguistic functions, either marking prominence or information structure, or marking prosodic structure, or both.

Though changes in pitch heavily contribute to the perception of prominence and the degree of juncture, perception of prominence and juncture is influenced by multiple sources of information such as acoustic differences in F0, duration and intensity, voice quality, and even linguistic structures (Cole, Mo, and Baek 2010; Cole, Mo, and Hasegawa-Johnson 2010; Bishop 2012). Because perception can be subjectively influenced by an individual's bias and linguistic background, ToBI requires observation of visual representations of acoustics, especially the F0 track and the waveform/spectrogram of the utterance. This holistic approach to transcription is necessary because ToBI transcription is phonological, capturing the information of metrical and prosodic structure as well as the tonal category. In this way, it is different from other transcription systems focusing on only F0 representations, such as INTSINT (International Transcription System for Intonation; Hirst 1998, and chapter 3, this volume; Hirst and Di Cristo 1998). In INTSINT, an F0 contour is labeled as a sequence of symbols referring to either an absolute tone level (e.g., *t* or ⇑ for the top of the speaker's pitch range) or a relative tone (e.g., *h* or ↑ for a higher tone than the preceding tonal segment), without considering the function of the tone or the prosodic structure of an utterance. Thus, the ToBI tone labels can be equated to broad phonemic transcription of the International Phonetic Alphabet (IPA), while INTSINT symbols correspond to a narrow phonetic transcription of the IPA. That is, the goal of ToBI is not simply to regenerate the surface F0 contours of any sentence in any language, but to capture tonal properties that are linguistically meaningful, especially for the function of marking prominence and prosodic structure.

Because ToBI is language-specific and the tonal categories of each ToBI system are contrastive in each language, defined based on the same theoretical assumptions and principles and using a common vocabulary, comparing the tonal types and the structure of tonal sequences across various ToBI systems allows us to identify intonational universals and study prosodic typology.

For example, as mentioned in sections 4.2 and 4.3.2, the function of ToBI tones, whether they mark prominence or a prosodic structure, can be identified by tonal symbols, especially diacritics. In the ToBI system of languages that have stress, a prominence marking tone is labeled with a diacritic * (e.g., H*, L*+H), called a (postlexical) pitch accent, while tones marking the boundary of a prosodic unit are labeled with an edge-marking diacritic such as %, -, or *a* (to mark a boundary tone of an IP, ip, and AP, respectively). But the * diacritic is also added to a lexical pitch accent, as shown in the Japanese and Serbo-Croatian ToBI systems. Unlike the postlexical pitch accent found in English, the lexical pitch accent in Japanese is not metrically strong, so an accented mora is not realized with higher amplitude or longer duration than an unaccented one (Beckman 1986). Thus, a starred tone marks a syllable or mora's special status as the "head" of a word, but that head is not necessarily associated with stress. If a language has no lexical prosody and has no word head, as in Korean, Mongolian, and West Greenlandic, no starred tone is used to categorize the F0 contour. In these headless languages, distinctive tonal categories are all phrasal tones or boundary tones marking a prosodic unit (e.g., an HLH tonal rhythm in West Greenlandic aligned with the right edge of a phonological word [Arnhold 2014] or a rising boundary tone on the left edge of an intermediate phrase [i.e., -LH] in Mongolian [Karlsson 2014]). In these headless languages, a word becomes prominent by being located at the edge of a larger prosodic unit. In Jun's (2014a) model of prosodic typology, these languages are categorized as edge-prominence languages, while languages like English, where prominence is marked by the head of a prosodic unit, are categorized as head-prominence languages (cf. Beckman 1996).[12] So, comparing diacritics used in the tonal labels of a ToBI system can provide data for typological analyses of prominence marking and prosodic structure.

Along the same line, by comparing tonal categories across ToBI systems, we can identify the most common versus rare tonal categories, or the most common prosodic units across languages. We can also learn which languages have a contrast on the rising slopes (e.g., H* versus L+H*) or on differences in the alignment of rises (e.g., L+H* versus L*+H) or falls (e.g., H+L*, H*+L), and so on. Languages can also be classified based on the number and type of pitch accents and boundary tones they have, and on how regular their tonal pattern is (for such comparisons, see Jun 2005b, 2014a).

Labeling distinctive categories of prosody by referring to their phonetic realizations in ToBI can also provide excellent data for studying the acquisition of prosody as a second or foreign language or the prosody of languages in contact. This is because one needs to know the distinctive tonal categories and prosodic structure, as well as their phonetic realizations, in both languages to study prosodic transfer or interference between them (e.g., McGory 1997; Ueyama and Jun 1998; Jun and Oh 2000; Dupoux et al. 2008; Atterer and Ladd 2004; Mennen 2004, 2007, 2015; Schepman, Lickley, and Ladd 2006; Trofimovich and Baker 2006; O'Brien and Gut 2010; Lee and Jun 2016). For example, both Japanese and Korean L2 (second language) speakers of English produced the nuclear pitch accent on the focused word in English declarative sentences (i.e., L+H*) better than that in English interrogative sentences (L*). This is because a focused word in Japanese and Korean is always produced with an H tone regardless of the sentence type (McGory 1997; Ueyama and Jun 1998). This suggests that difficulty in producing L2 prosody is due to the differences in the prosodic category between L2 and the speaker's native language (L1). Difficulty in producing L2 prosody can also come from the differences in phonetic realizations of the same prosodic category between L1 and L2. An example can be seen when bilingual speakers of German tend to produce a delayed onset of F0 rise when producing English prenuclear rising pitch accent, because the onset of F0 rise is aligned later in German than in English (Atterer and Ladd 2004). Similarly, both Greek and Dutch have a prenuclear rising pitch accent, but Dutch speakers of Greek often produce an earlier F0 peak when producing Greek prenuclear rising pitch accent, because the F0 peak is aligned later than the accented syllable in Greek than in Dutch (Mennen 1998, 2004). Transcribing the prosody of utterances produced by L2 speakers or bilinguals by referring to the ToBI systems of the two languages involved reveals how the prosodic properties of two languages, either categorical or phonetic, interact in the speech of L2 learners or bilinguals (Lee and Jun 2016).

An additional strength of the ToBI system can come from the flexibility of adding new tiers. Each tier in ToBI can be used to transcribe different prosodic and linguistic information, and the misc tier can even carry nonlinguistic information. As noted in section 4.3.6, ToBI systems for various languages have added new tiers to tag language-specific prosodic information as well as other linguistic information tailored to the particular interests of the research community of the ToBI variety. For example, C-ToBI added a foot tier to transcribe the fusion forms, that is, to mark the degree of lenition between two adjacent syllables, as well as emphasized syllables and phrases. The goal of transcribing syllable fusion was to explore the factors affecting syllable fusion, and the goal of transcribing emphasized syllables was to explore whether the domain of prominence has any relationship with the prosodic hierarchy.[13] The addition of a phonetic tone tier in K-ToBI also allowed researchers to explore the factors affecting the various tonal patterns observed in the AP. By transcribing the allotonic surface tonal categories in the phonetic tone tier for various utterances produced in different speech styles, Yoo and Jun (2016) recently found that, between the two AP-medial tones, the H tone is realized more frequently than the L tone and is more common in spontaneous speech than in formal speech. A new tier can therefore be added to a ToBI system of a specific language for an exploratory purpose, and what to label in the tier can be decided by the particular group of researchers working on the target language variety.

In addition to testing and improving existing models of intonation by transcribing language-specific information in a new tier, ToBI also represents a useful tool for studying the interfaces between prosody and various subareas of linguistics. Utilizing parallel quasi-independent multitiers aligned with each other, researchers can create and examine ToBI-labeled corpora (e.g, Ostendorf et al. 2001; Grice and Savino 2003) or add ToBI labels to an existing database such as the English Broadcast News Speech (Fiscus et al. 1998) and the Buckeye Corpus (Pitt et al. 2005). For example, the MAE_ToBI-labeled corpus of the Boston University Radio Speech Corpus (Ostendorf, Price, and Shattuck-Hufnagel 1995) was examined to study the realization of glottal stops and glottalization of word-initial vowels in different prosodic positions (e.g., Dilley, Shattuck-Hufnagel, and Ostendorf 1996; Garellek 2013). News speech corpus (Fiscus et al. 1998) and National Public Radio news excerpts have also been labeled in MAE_ToBI to examine the prosodic realization of reflexive pronouns (e.g., *himself, themselves*) relative to their syntactic and semantic structures (Ahn 2015). Similarly, Grice and Savino (2003) analyzed dialogues from the map tasks in Bari Italian by adopting the coding scheme used in the English Human Communication Research Center (HCRC) Map Task (Anderson et al. 1991) and transcribed the intonation of the utterances in the corpus following the ToBI notation in Grice et al. (2005). They discovered that in Bari Italian, polar questions asking about new information employ a rising pitch accent (L+H*), while questions about given information (i.e., confirmation-seeking questions) are expressed with a falling pitch accent (H*+L or H+L* depending on a contrast setting). Furthermore, Stirling et al. (2001) transcribed part of the speech corpora from the Australian National Database of Spoken Language (ANDOSL) map task corpus (Miller et al. 1994) in Australian English ToBI (AuE_ToBI, Fletcher and Harrington 1996) and investigated correlations between discourse structure and various prosodic properties.

Finally, researchers can also use tonal categories and BIs from ToBI transcriptions to test linguistic theories and hypotheses in experimental contexts. ToBI is a particularly useful tool when manipulating prosody for experimental stimuli and for evaluating and categorizing prosodic structures produced by experimental subjects. Perhaps the best illustration of this type of use is research on the role of prosody in sentence processing (e.g., Speer, Kjelgaard, and Dobroth 1996; Schafer et al. 1996; Kjelgaard and Speer 1999; Schafer et al. 2000; Schafer, Speer, and Warren 2005; Ito and Speer 2008; Snedeker and Casserly 2010; Jun 2010; Lee and Watson 2011; Speer and Foltz 2015; Jun and Bishop 2015). These studies explored how manipulations of prosodic structure influence listeners' interpretation and processing of syntactic structure, or, conversely, how their production of prosody changes as a result of manipulating syntactic structure or pragmatic context. Prosodic manipulation and evaluation in these studies was facilitated by the adoption of the tonal categories and/or BIs from the ToBI conventions of the target language. For example, to test Fodor's implicit prosody hypothesis (Fodor 1998, 2002),[14] the prosodic phrasing of sentences produced by subjects (e.g., "Someone shot the servant of the actress who was on the balcony") was assessed by labeling a BI and a boundary tone, if present, after each head noun ("the servant" and "the actress"; see Bergmann, Armstrong, and Maday 2008 and Jun 2010 for English data using MAE_ToBI; Jun and Koike 2008 for Japanese data using J_ToBI; and Jun and

Kim 2004 for Korean data using K-ToBI). Speer and Foltz (2015) also tested the implicit prosody hypothesis by manipulating the presence and type of pitch accent on target words in visual-to-auditory cross-modal priming experiments. Based on ToBI-labeling speech samples taken from paragraphs read aloud, they found that only speakers who reliably produced L+H* pitch accents on contrastively focused words in the produced speech sample (i.e., with explicit prosody) responded faster to an auditory probe word in L+H* following the silent reading of a correctively focused prime word. That is, priming of contrastive intonation from implicit to explicit prosody depended on the speech styles of individual speakers, which were established by way of a ToBI annotation of individual speakers' prosody.

In sum, the ToBI system, based on the AM model of intonational phonology, is an important tool for advancing knowledge on the phonetics and phonology of prosody. It is particularly useful for illuminating the relation between prosody and other linguistic phenomena in a way not possible before the system's creation. As mentioned in Beckman, Hirschberg, and Shattuck-Hufnagel (2005, 46), ToBI is "an ongoing research program rather than a set of rules cast in stone for all time." In addition to facilitating linguistic research, it is a valuable tool for creating communal corpora that can be used to address a variety of research questions. It can also be used by engineers and speech scientists to enhance the performance of systems for recognizing and synthesizing intonation. With ToBI-labeled data, we can investigate how prosodic categories are phonetically realized, and how prosody delivers phonological, syntactic, semantic, and discourse-related information. We can also explore how languages differ in their marking of such complex information prosodically.

4.5 Challenges and Recent Developments

This section describes challenges that ToBI users face as the result of some of the system's weaknesses. While some of these challenges stem from properties of the AM theory that ToBI adopts, others arise from the nature of transcribing real speech using a discrete phonological transcription system. This section also describes some of the recent developments the ToBI community has devised to address ToBI's weaknesses.

One of the challenges ToBI users encounter is that the prominence patterns and prosodic structure captured by ToBI transcriptions are limited to the size of an utterance corresponding to an IP. This is a consequence of the fact that, in the AM model of intonational phonology, the IP is the highest prosodic unit defined by a tone.[15] Therefore, prosodic events that occur across an IP boundary or over a sequence of IPs cannot be captured by the current ToBI system. This makes it difficult to use ToBI to capture how prosody is used to mark complex discourse structure, or possible syntagmatic relations between tones across IPs (e.g., Ladd 1990, [1996] 2008a; Calhoun 2006, 2012; Bishop 2013).[16] One of the main prosodic cues known to mark discourse structure is pitch range (Brown, Currie, and Kenworthy 1980; Hirschberg and Pierrehumbert 1986). While ToBI includes pitch reset among the acoustic cues defining a prosodic unit, no labels exist to cover pitch range changes across ips and IPs. Mandarin ToBI (Peng et al. 2005) includes a limited number of labels for pitch range, but they refer to effects within one sentence or phrase (i.e., one breath group): that is, %reset to mark pitch reset over a declarative phrase, %q-raise to mark raised pitch range (as in echo questions), %e-prom for the local expansion of pitch range within a phrase due to emphatic prominence, and %compressed for the local compression of pitch range after %e-prom. The MAE_ToBI system does include the label HiF0, intended to mark the highest F0 point for a pitch accent within an ip. But this was simply intended as a tag for accessing the F0 value for use later as a measure of pitch range to study the relationship between pitch range and discourse structure. Instead of using a tones tier, a new tier could be added for pitch-range information. J_ToBI, for example, added an extra tier to mark finality. Though the finality tier for J_ToBI only marks a juncture larger than an IP (delivering "the sense that the speaker is 'done' in her turn in discourse planning"; Venditti 1997, 2005), this type of tier could be used to label acoustic cues to the relationship between IPs and larger discourse segments.

Because most of the existing ToBI systems were developed on the basis of sentence-length laboratory speech with relatively simple discourse structure (e.g., reading of a story, news reports, interviews, short dialogues, speech from map tasks), ToBI transcriptions have not been widely used by researchers working on the discourse structure of free conversational speech. However, as noted in Beckman, Hirschberg, and Shattuck-Hufnagel (2005, 12), transcription conventions for ToBI are ideally "based on a large and long-established body of research on dialectology, pragmatics and discourse analysis for the target language," in addition to study of its intonational phonology. The inclusion of a tier that allows for annotating relationships between ips and IPs (or between larger discourse segments) would thus be a straightforward way to address ToBI's current weakness in this area.

In a related vein, the current ToBI system is also not optimized for labeling global, and gradual, prosodic events. This stems from basic properties of the AM model of intonation, which represents the intonation contour as a linear sequence of low and high tonal targets. If a researcher is more interested in the domain of global prosodic events (e.g., pitch range reset at the beginning of prosodic unit X or Y) or relative differences between domains (e.g., whether the pitch range of unit X is larger or smaller than that of unit Y), this could be addressed by creating a label to mark the onset or magnitude of the prosodic feature at the edges of units. However, modifications that require a researcher to address more gradual changes over domains present a much greater challenge to the ToBI system (e.g., annotating changes in the slope of declination, which marks sentence types in Danish intonation; Thorsen 1980, 1983).

Given the variability and gradience inherent in acoustic signals, the next challenge ToBI users face is not due to the system's reliance on the AM model specifically but to its basic status as phonological. First, prosodic categories are not always realized in the same way. As Cole and Shattuck-Hufnagel (2016), Cangemi and Grice (2016), and Arvaniti (2016) have emphasized, acoustic cues to a prosodic category are not always realized consistently. Instead, factors such as time pressures and segmental/prosodic context often render cues reduced or ambiguous, and prosodic categories can also be realized differently depending on the individual speaker's strategy to encoding prosodic contrasts (e.g., Mo, Cole, and Lee 2008; Cole and Shattuck-Hufnagel 2016). Moreover, criteria for defining prosodic categories, such as phrase-final lengthening for an IP or ip boundary, are gradient rather than categorical. Categorizing a prosodic event based on gradient and variable acoustic cues is not straightforward in the absence of other information supporting the category, and the final decision can be further influenced by the individual labeler's weighting of the acoustic cues.

Second, and related to this, recall from section 4.4 how ToBI labels are decided based on the labeler's perception of prominence and degree of juncture from an audio file, while at the same time observing the visual display of a pitch track together with a spectrogram or waveform. This means that the labeling is not fully based on objective criteria and can be influenced by the labeler's perception of acoustic cues, interpretation of linguistic context, and level of knowledge and experience in ToBI labeling. It is therefore recommended that ToBI labeling be done by multiple trained labelers to reduce the subjectivity of the annotations. However, because ToBI labeling is labor intensive and requires training in both acoustics and AM theory, it is not always easy to find multiple ToBI labelers to participate in the annotation of experimental data or databases.

The fact that ToBI systems are tools for phonological transcription means they are well suited to studying prosodic typology. One of the strengths of a phonological transcription system is that it enables comparison of how different languages (and language varieties) tonally mark prominence and prosodic structure. However, this also comes with a challenge. Because ToBI's tone labels are language-specific, reflecting contrasts in each language, the same surface F0 contour is not necessarily assigned the same labels across languages (just as the same Voice Onset Time [VOT] can be categorized as /p/ in Spanish but as /b/ in English). Additionally, as pointed out by Ladd (2008b) and Hualde and Prieto (2016), it is also possible that the same surface F0 contours, representing the same contrast in each language, can be labeled differently depending on what analysis and tonal categories are chosen by the developers of the ToBI system.

A good example illustrating this point can be found in the comparison of MAE_ToBI and GToBI (Grice et al. 2005). Both American English and Standard German have four levels of sentence-final pitch, representing the same basic meaning/function: low F0 at the end of a neutral statement, mid F0 at the end of a calling contour, high flat F0 for incompleteness, and super-high F0 at the end of a yes-no question. However, these four types of contours are transcribed differently: as L-L%, !H-L%, H-L%, and H-H% in MAE_ToBI, but as L-%, !H-%, H-%, and H-^H% in GToBI. This is because English ToBI developers adopted an upstep rule (i.e., an IP boundary tone is realized higher after a (!)H- phrase accent), but GToBI developers did not. Instead, GToBI's creators added two symbols intended to make their system for German more phonetically transparent. One symbol, an upstep diacritic ^, is used to directly mark a super-high tone, ^H; another, a toneless IP boundary symbol %, is used when the IP-final pitch level is the same as the ip-final pitch (i.e., H-% means a high plateau and L-% a low plateau). Thus the difference between these two ToBI systems was a matter of choosing how abstract versus how phonetically transparent the tonal representation should be. Because the tones in the original ToBI system, by introducing a downstep diacritic, were claimed to be more faithful to the surface F0 patterns than those proposed in the original AM model (see section 4.3.2), various ToBI systems have added tones and symbols to better capture surface F0 contours, but there are still no widely agreed-on rules or conventions regarding what symbols and labels to use and on what occasions to use them. (See the "portability" problem of ToBI notation argued in Hualde and Prieto 2016.) As more ToBI systems are developed in the future, there will be new symbols and labels created, making comparison across ToBI systems more difficult and the study of prosodic typology thus more challenging. Students and researchers will also likely find it more difficult to determine what some labels and symbols in different ToBI systems mean[17] and what level of abstraction they represent.[18]

In addition to the difficulty in interpreting symbols and levels of abstraction in different ToBI systems generally, there is also the challenge of deciding what level of abstraction is appropriate when a surface tone does not match an underlying tone. Should it be transcribed with an allophonic tone label or with an underlying tone label? As a system designed to transcribe phonological categories, clearly the underlying tone should be used. However, if the goal is simply to tag one's observations about the speech signal, one might want to use an allophonic tone label. Cases like this have been a known source of confusion and disagreement among labelers (e.g., Escudero et al. 2012; Armstrong 2017; Hualde and Prieto 2016), and the chosen solutions vary across ToBI systems. For example, when an underlying L% boundary tone is truncated after an L+H* on a phrase-final syllable (and thus realized as an approximately mid-level tone), should the boundary tone be labeled as L%, !H%, or something else? Prieto and Ortega-Liebaria (2009) and Hualde and Prieto (2016) used !H% in Catalan and Peninsular Spanish ToBIs, but Grice et al. (2005) used (L%)—L% in parentheses—for southern Italian varieties, as a notation for a partially realized underlying tone.

As a tagging tool, the ToBI system can include labels that represent surface tonal patterns, and this makes its level of tonal representation clearly less abstract than that proposed in the AM model of intonational phonology. However, the common practice of using ToBI-style labeling (employing primarily the words and tones tiers) when trying to build a model of intonational phonology of a language seems to have blurred the distinction between the ToBI system and the AM model of intonational phonology. This trend can be observed in publications such as Prieto and Roseano (2010), Jun (2014b), and Frota and Prieto (2015), where most models of intonational phonology and most ToBI transcription systems look quite similar to each other. Consequently, the BI part of ToBI has also become less emphasized in the ToBI system. In fact, it is not uncommon to hear researchers refer to the "ToBI model" of a particular language, which would seem to equate the transcription system with the intonational phonology model of the language.[19]

Many of the issues and challenges described herein motivated a recent proposal to develop an international prosodic alphabet (IPrA) within the AM framework (Hualde and Prieto 2016 and Jun and Fletcher 2014; see also presentations by Jun, Prieto, and Hualde at the satellite workshop of the International Congress of Phonetic Sciences 2015 in Glasgow[20]). The current proposal calls for ToBI systems to include two levels of tonal transcription—one distinctive, thus phonological, and the other nondistinctive but categorical, thus phonetic—and to create a set of discrete IPrA tonal labels and diacritics (the details of which are to be agreed on by the international community of ToBI users). The goal is that the symbols in IPrA can be used for both phonological and categorical phonetic transcriptions, like the symbols in the IPA, so that they can be used in both phonemic and phonetic representations. The labels for *phonological* transcription would only mark phonological contrast in the language, and the *categorical phonetic* labels would be for tones that are categorical in nature but not distinctive, or whose distinctiveness is not yet known. These categorical phonetic labels can (i) represent allophonic realizations of an underlying tonal category, (ii) be used as temporary labels before their phonological status is established (e.g., at the beginning stage of building a model of intonational phonology), or (iii) represent hybrid or exceptional tonal categories that are not part of the intonational model of any specific language (e.g., L2 speech, bilinguals' speech, or speech produced by speakers of multiple languages in contact). The phonological tone labels can be used in the tones tier, and categorical phonetic tones can be labeled in a new, separate tier (e.g., an IPrA tier) or can be used in the tones tier together with the phonological tones, but written in square brackets (e.g., [L*] for an allophonic L*). It is hoped that separating categorical phonetic tones from phonological tones in ToBI will facilitate the development of more abstract phonological analyses of intonation and better represent the correspondence between underlying tonal categories and surface patterns in a systematic way. This will help us to clarify the levels of transcription that different ToBI systems are using and make more transparent comparisons across languages possible. Clarifying the levels of transcription will also increase interlabeler agreement rates and facilitate labeling large corpora and further improve the strengths of the ToBI system mentioned in section 4.4.

In addition to the IPrA proposal, some of the weaknesses of the ToBI system motivated the recent development of a few other transcription systems such as rapid prosody transcription (RPT) and rhythm and pitch (RaP) transcription. The RPT system was developed by Cole and her colleagues (Mo, Cole, and Lee 2008; Cole, Mo, and Baek 2010; and Cole, Mahrt, and Hualde 2014). Unlike ToBI transcription, which is labor-intensive and requires training and knowledge of the theoretical background, the RPT transcription is, as the name implies, a coarse-grained transcription done by untrained naive native speakers of the target language.[21] Because this type of annotation can be done as a group, it is possible to generate transcriptions from many labelers in a very short time. In the RPT task, untrained native speakers (such as college students) are given instructions defining "prominent" words and "prosodic juncture" in running speech. They are then asked to listen to a sample of running speech and identify prominent words and prosodic boundaries on a printed or electronically presented transcript of the speech sample. Identification of these prosodic events is done in real time (i.e., the "rapid" part of RPT), with the subjects typically listening to the entire speech sample four times, twice each for prominence and boundary transcription (they are not allowed to replay any specific parts of the sample). The length of speech samples used can vary from ten to sixty seconds, which generally includes one or more syntactic clauses and one or more prosodic phrases. In Cole and colleagues' studies, the task was used to generate a prominence score (p-score) and a boundary score (b-score) for each word (i.e., the proportion of labelers who marked each word as prominent or as preceding a boundary, respectively). The magnitude of p-scores and b-scores for individual words and their distribution thus provides a measure of agreement among labelers and an estimation of the variability in the perception of prosodic events in the speech material (Mo, Cole, and Lee 2008; Cole and Shattuck-Hufnagel 2016).

Bishop and his colleagues (Bishop and Kuo 2016; Bishop, Kuo, and Kim 2020) compared the responses from an RPT task with the labels provided by expert ToBI annotators. Their results showed that RPT subjects identified just under 50% of the nuclear pitch accents (NPAs) and just under 50% of the IP boundaries that ToBI labelers assigned to the speech materials. Thus a little more than half of the time, the naive labelers failed to identify NPA-level prominence and IP-level boundaries. This is not surprising given that the RPT labelers were not trained to attend to prominence and phrasing at a detailed level, and the transcription was done in real time without examining pitch tracks or being able to hear parts of the sentence as many times as desired. However, from the words identified as prominent or having a prosodic boundary by the RPT labelers, pitch-accented words were identified as prominent more often (i.e., had higher p-scores) than unaccented words, and words at an IP boundary were identified as a boundary more often (i.e., had higher b-scores) than those at an ip boundary. This suggests that the p- and b-scores from a large set of speech materials annotated by a large group of RPT labelers can provide a good measure of the perceptual salience of prosodic categories. RPT may also provide useful information about the acoustic correlates of various prosodic categories, as well as about how the perception of prominence and boundaries is influenced by grammatical information. As pointed out in Cole and Shattuck-Hufnagel (2016), the patterning of the p- and b-scores from RPT may shed light on the mechanisms involved in prosodic processing in speech production and perception.

Next, the RaP system was developed by Dilley and her colleagues (Dilley 2005; Dilley and Brown 2005; Dilley et al. 2006; Breen et al. 2012) to address transcription difficulties associated with ToBI's vulnerability to the variability and gradience of tonal categories and to reflect the importance of rhythmic or metrical prominence distinct from pitch accent. In ToBI, tonal labels such as H* or L-H% include both the identity of the tonal target (e.g., H or L or !H) and information regarding its function in terms of both metrical prominence (i.e., pitch accents are explicitly marked by *) and prosodic structure (i.e., boundary tones are marked by % and -). The RaP system, in contrast, emphasizes the independence of pitch information from rhythmic/metrical and phrasal structure. Pitch information is transcribed in terms of three tonal targets (H, L, E) whose value is relative to a preceding tone (higher, lower, or equal to the preceding tonal labels, respectively). Tones in RaP, labeled in the pitch tier, are therefore represented phonetically. Metrical prominence (in three levels: strong, weak, none)[22] and prosodic structure (in two levels: IP and ip) are transcribed in the rhythm tier. Although RaP was proposed as "a method of labeling the rhythm and relative pitch of spoken English" (Dilley and Brown 2005, 2), the concept and principles of the system can be applied to other languages, and the system can be used by researchers interested in examining structural information separately from tonal information.

4.6 Conclusion

This chapter introduced the ToBI system for transcribing phonological aspects of prosody. It discussed both the strengths and challenges associated with the system, as well as some of the recent developments the research community has made in response to such challenges. As a phonological annotation system based on the AM model of intonational phonology, each language's ToBI system is unique, reflecting the prosodic properties of the language in question. The ToBI framework allows a common set of conventions to be used across languages, despite the fact that how prominence and phrasal structure are phonologically marked and phonetically realized is language-specific. In the ToBI system, tones and BIs are annotated based on the labeler's perceived prominence of words and degree of juncture between words, supplemented by the visual display of the utterance's acoustics, especially F0 track and waveforms (or spectrogram). Furthermore, as a tool to annotate and communicate observations made about the speech signal with other users of ToBI, the ToBI system allows communal corpora to be created for testing hypotheses about various linguistic and prosodic phenomena. The output of such research also serves as feedback to the ToBI transcription system itself and informs models of intonational phonology.

It was shown that most of the strengths as well as most of the weaknesses associated with ToBI stem from its being a phonological transcription system based on the AM model. The ToBI system's emphasis on transcribing distinctive phonological categories of tones and metrical/prosodic structure makes it the most valuable tool currently available for studying the grammar of prosody and intonation. ToBI users can pursue questions about prosody's relation to a number of other linguistic subfields, and it is especially well-suited to the study of prosodic typology and universals. However, as a phonological system, the labels proposed in various ToBI systems do not necessarily have the same phonetic values or represent the same level of abstraction from system to system, and so an effort should be made when comparing multiple ToBI systems to study prosodic typology. Efforts to address this particular issue are currently underway, such as the recent proposal to develop an IPrA, allowing for the categorical phonetic representation of tones in each ToBI system.

ToBI is a valuable tool for conducting research on prosody, and it is a powerful and flexible transcription system. With approximately fifty ToBI systems now established from typologically different languages, tremendous progress has been made since the original ToBI system was published in 1994, in terms of both prosody's relation to other linguistic subfields and our understanding of prosodic typology. Although ToBI transcription is labor-intensive, and becoming a confident labeler requires significant training, the benefits seem clear and justified. Ongoing and future tasks for the ToBI research community include improving on the system's current limitations, as well as continuing to expand the number (and diversity) of languages described within the ToBI framework. Taken together, the ToBI framework represents a powerful and flexible tool, advancing our understanding of prosodic systems and prosodic typology.

Notes

1. Most of the ToBI systems presented at the meeting, including the MAE_ToBI, were later published in Jun (2005c).

2. As will be shown in sections 4.3.2 and 4.4, not all languages mark intonational prominence by way of pitch accent. If a language has no stress or lexically marked syllables—and thus has no designated terminal element for words (e.g., Korean, Mongolian, or West Greenlandic), prominence is not marked by a pitch accent but by positioning the prominent word at the edge of a prosodic unit. Thus how prosodic prominence is marked varies across languages. See Jun (2014a) for a typology of prominence marking.

3. This prosodic structure is similar to, but slightly different from, the prosodic hierarchy proposed by prosodic phonologists in the 1980s, which was defined indirectly based on the syntactic structure of a sentence (e.g., Nespor and Vogel 1986; Selkirk 1986; Hayes 1989). In that approach, prosodic units commonly assumed to be larger than a word include, from the largest to the smallest, an utterance (Utt) > an intonational phrase (IPh) > a phonological phrase (PPh) > clitic group. Both an Utt and an IPh correspond to an IP in intonational phonology, but a PPh corresponds to either an ip or an AP in intonational phonology (see Jun 1998 for comparisons between these two approaches).

4. The H*+L pitch accent proposed in Pierrehumbert (1980) and Beckman and Pierrehumbert (1986) is not a falling tone as implied by the trailing L tone. As mentioned in Beckman, Hirschberg, and Shattuck-Hufnagel (2005), it is for a simple H* pitch accent, followed by a downstepped H*. The trailing L was simply included to provide the necessary bitonality required, as shown in the research on African tone languages (e.g., Leben 1973; Goldsmith 1976), to trigger downstep of a following H tone. That is, a sequence of H*+L pitch accents was used to refer to a series of stepping-down pitch accents, examined in Liberman and Pierrehumbert (1984).

5. Allowing tonal labels to be more faithful to the surface F0 contour became the most prominent difference between the tonal inventory of the AM model of intonational phonology and that of the ToBI system. As will be discussed in section 4.5, this difference was a source of confusion to many researchers and significantly affected subsequent models of the intonational phonology of other languages.

6. A prosodic word can also be marked by intonation as in Serbo-Croatian (Godjevac 2005).

7. Because the undershoot allophones of an L boundary tone are predictable from the words tier in J_ToBI, both wL% and %wL symbols are removed in X-JToBI (Maekawa et al. 2002), a revised version of J_ToBI.

8. If the AP-initial segment is either an aspirated or a tense consonant, the tone is H. Otherwise, the tone is L.

9. The GToBI had the same number of prosodic units as in MAE_ToBI, but the GToBI developers proposed not to label a BI lower than 3 and proposed to label two types of mismatch cases, labeled 2r and 2t. BI 2r refers to rhythmic break with tonal continuity, and BI 2t refers to tonal break with rhythmic continuity.

10. In the original ToBI, filler words (such as *uh* and *um*, also known as filled pauses) were not treated as a type of disfluency. But this and other types of disfluency could be added in other ToBI systems.

11. When a labeler is sure about the degree of juncture at a certain word boundary but the tonal realization does not match the corresponding degree of juncture, a mismatch label (2 or #m, discussed in section 4.3.3) should be used.

12. Languages can cue a word's prominence by marking both the head and the edge of the word, and this type of language is categorized as a head-/edge-prominence language (examples are Georgian, Bengali, French, and Japanese). The tonal inventory of these languages includes a starred tone as well as a boundary tone of a word or AP (see Jun 2014a).

13. In C-ToBI, the domain of the foot is reflected in the BI tier by aligning the foot-medial weakened syllable with a BI 0, while the foot-final syllable is aligned with a BI 1. However, the domain of prominence is not linked with the BI tier, reflecting the unknown relationship between the domain of prominence and the prosodic hierarchy. The developers of C-ToBI state that the architecture of independent parallel tiers "highlights one of the principal advantages of the looser structure of the original ToBI framework" (Wong, Chan, and Beckman 2005, 293).

14. The hypothesis was proposed to explain cross-linguistic differences in attachment preferences. Fodor claimed that the difference is due to implicit default prosody assigned to the syntactic structure of each sentence. She predicted that, in the example sentence where the attachment of a relative clause (RC) is ambiguous between the first noun ("the servant") or the second ("the actress"), languages that prefer high attachment (i.e., RC attachment to the first noun) would have a larger juncture after the second noun than after the first noun, and the opposite is true for low attachment preference languages. She also assumed implicit prosody generated in silent reading is equal to explicit prosody generated in reading aloud.

15. The AM model does not include a prosodic unit above the IP, such as the utterance-level unit proposed by some prosodic phonologists (e.g., Nespor and Vogel 1986). The primary reason for this is the absence of distinctive tones or other categorical prosodic events that mark any unit above the IP.

16. In the AM model of English intonation adopted in the MAE_ToBI, an ip is the domain of nuclear pitch accent and the domain of pitch reset. Thus, capturing prominence relation across ips is limited.

17. Examples where ToBI systems assign different symbols for the same phenomenon are easily found among the systems currently available. For example, undershoot has been marked with either *w* (in Japanese and Greek ToBIs) or *u* (in Bangladesh Bengali ToBI), and an upstep has been marked with ^ (in German and Bininj Gun-wok ToBI) or with ¡ (in Spanish and Catalan ToBIs).

18. For example, downstep can be phonological or phonetic across languages. Its status can also vary within a language depending on tone types (e.g., in GRToBI, downstepped H is contrastive for a phrase accent (!H-) or a boundary tone (!H%) but not for a pitch accent (e.g., !H* or !H*+L).

19. In this case, a “ToBI model” would seem to refer to an intonational phonology model that is more transparent and surface-rich than the traditional AM model of intonational phonology.

20. For more information about the workshop, visit http://linguistics.ucla.edu/ipra_workshop/.

21. The RPT methods have been used for Hindi, Russian, French, and Spanish in addition to English (for some cautionary notes on applying the RPT across languages, however, see Gussenhoven 2015). It has also been used with fluent bilinguals and learners of English as L2 (see Cole and Shattuck-Hufnagel 2016).

22. In English ToBI, metrical prominence is also represented at three levels (Beckman & Edwards 1994): NPA, pitch accent, no accent. The NPA is the last pitch accent within an ip. Because it is predictable from the location of pitch accent and an ip boundary tone, NPA is not labeled with a separate symbol in English ToBI. But in languages where NPA is not predictable as in Italian, an NPA tone is marked by a diacritic *n* (e.g., L* + Hn).

References

Ahn, B. 2015. “Giving Reflexivity a Voice: Twin Reflexives in English.” PhD diss., UCLA.

Anderson, A., M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, et al. 1991. “The HCRC Map Task Corpus.” *Language and Speech* 34:351–366.

Arbisi-Kelm, T. 2006. “An Intonational Analysis of Disfluency Patterns in the Speech of Stutterers.” PhD diss., UCLA.

Armstrong, M. E. 2017. “Accounting for Intonational Form and Function in Puerto Rican Spanish Polar Questions.” *Probus* 29 (1): 1–40. https://doi.org/10.1515/probus-2014-0016.

Arnhold, A. 2014. “Prosodic Structure and Focus Realization in West Greenlandic.” In Jun 2014b, 216–251.

Arvaniti, A. 2016. “Analytical Decisions in Intonation Research and the Role of Representations: Lessons from Romani.” *Laboratory Phonology* 7 (1): 1–43.

Arvaniti, A., and M. Baltazani. 2005. “Intonational Analysis and Prosodic Annotation of Greek Spoken Corpora.” In Jun 2005c, 84–117.

Arvaniti, A., D. R Ladd, and I. Mennen. 1998. “Stability of Tonal Alignment: The Case of Greek Prenuclear Accents.” *Journal of Phonetics* 26:3–25.

Arvaniti, A., D. R. Ladd, and I. Mennen. 2000. “What Is a Starred Tone? Evidence from Greek.” In *Papers in Laboratory Phonology V*, edited by M. Broe and J. Pierrehumbert, 119–131. Cambridge: Cambridge University Press.

Atterer, M., and D. R. Ladd. 2004. “On the Phonetics and Phonology of “Segmental Anchoring” of F0: Evidence from German.” *Journal of Phonetics* 32: 177–197.

Beckman, M. E. 1986. *Stress and Non-Stress Accent*. Dordrecht, the Netherlands: Foris.

Beckman, M. E. 1996. “The Parsing of Prosody.” *Language and Cognitive Processes* 11:17–67.

Beckman, M. E., and G. M. Ayers Elam. 1997. “Guidelines for ToBI Labeling.” Unpublished manuscript. Ohio State University. https://www.ling.ohio-state.edu/research/phonetics/E_ToBI/.

Beckman, M. E., M. Díaz-Campos, J. T. McGory, and T. A. Morgan, 2002. “Intonation across Spanish, in the Tones and Break Indices Framework.” *Probus* 14:9–36.

Beckman, M. E., and J. Edwards. 1994. "Articulatory Evidence for Differentiating Stress Categories." In *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, edited by P. A. Keating, 7–33. Cambridge: Cambridge University Press.

Beckman, M. E., and J. Hirschberg. 1994. "The ToBI Annotation Conventions." Unpublished manuscript. Ohio State University. https://web.archive.org/web/20100622195855/http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/ToBI/ToBI.6.html.

Beckman, M. E., J. Hirschberg, and S. Shattuck-Hufnagel. 2005. "The Original ToBI System and the Evolution of the ToBI Framework." In Jun 2005c, 9–54.

Beckman, M. E., and S.-A. Jun. 1996. "K-ToBI (Korean ToBI) Labelling Convention." Version 2. Unpublished manuscript. Ohio State University and UCLA. https://linguistics.ucla.edu/people/jun/ktobi/k-tobi-V2.html.

Beckman, M. E., and J. Pierrehumbert. 1986. "Intonational Structure in Japanese and English." *Phonology Yearbook* 3:255–309.

Bergmann, A., M. Armstrong, and K. Maday, K. 2008. "Relative Clause Attachment in English and Spanish: A Production Study." In *Proceedings of the Fourth International Conference on Speech Prosody*. https://www.isca-speech.org/archive/sp2008/papers/sp08_505.pdf.

Bishop, J. 2012. "Information Structural Expectations in the Perception of Prosodic Prominence." In *Prosody and Meaning (Interface Explorations)*, edited by G. Elordieta and P. Prieto, 239–269. Berlin: Mouton de Gruyter.

Bishop, J. 2013. "Prenuclear Prominence: Phonetics, Phonology, and Information Structure." PhD diss., UCLA.

Bishop, J., and J. Fletcher. 2005. "Intonation in Six Dialects of Bininj Gun-wok." In Jun 2005c, 331–361.

Bishop, J., and G. Kuo. 2016. "Do 'Autistic-Like' Personality Traits Predict Prosody Perception?" Paper presented at LabPhon 15 (Workshop on Personality in Speech), Cornell University, Ithaca, New York, July 17.

Bishop, J., G. Kuo, and B. Kim. 2020. "Phonology, Phonetics, and Signal-Extrinsic Factors in the Perception of Prosodic Prominence: Evidence from Rapid Prosody Transcription." *Journal of Phonetics* 82. https://doi.org/10.1016/j.wocn.2020.100977.

Bolinger, D. L. 1951. "Intonation: Levels versus Configurations." *Word* 7:199–210.

Bolinger, D. L. 1958. "A Theory of Pitch Accent in English." *Word* 14:109–149.

Breen, M., L. C. Dilley, J. Kraemer, and E. Gibson. 2012. "Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)." *Corpus Linguistics and Linguistic Theory* 8:277–312.

Brown, G., K. Currie, and J. Kenworthy 1980. *Questions of Intonation*. London: Croom-Helm.

Bruce, G. 1977. *Swedish Word Accents in Sentence Perspective*. Lund, Sweden: Gleerup.

Brugos, A., N. Veilleux, M. Breen, and S. Shattuck-Hufnagel. 2008. "The Alternatives (Alt) Tier for ToBI: Advantages of Capturing Prosodic Ambiguity." In *Proceedings of the Fourth International Conference on Speech Prosody*. https://www.isca-speech.org/archive/sp2008/papers/sp08_273.pdf.

Calhoun, S. 2006. "Information Structure and the Prosodic Structure of English: A Probabilistic Relationship." PhD diss., University of Edinburgh.

Calhoun, S. 2012. "The Theme/Rheme Distinction: Accent Type or Relative Prominence?" *Journal of Phonetics* 40:329–349.

Cangemi, F., and M. Grice. 2016. "The Importance of a Distributional Approach to Categoriality in Autosegmental-Metrical Accounts of Intonation." *Laboratory Phonology* 7 (1): 1–20.

Cole, J., T. Mahrt, and J. I. Hualde. 2014. "Listening for Sound, Listening for Meaning: Task Effects on Prosodic Effects on Prosodic Transcription." In *Proceedings of the Seventh International Conference on Speech Prosody*. https://www.isca-speech.org/archive/SpeechProsody_2014/pdfs/165.pdf.

Cole, J., Y. Mo, and S. Baek. 2010. "The Role of Syntactic Structure in Guiding Prosody Perception with Ordinary Listeners and Everyday Speech." *Language and Cognitive Processes* 25:1141–1177.

Cole, J., Y. Mo, and M. Hasegawa-Johnson. 2010. "Signal-Based and Expectation-Based Factors in the Perception of Prosodic Prominence." *Laboratory Phonology* 1:425–452.

Cole, J., and S. Shattuck-Hufnagel. 2016. "New Methods for Prosodic Transcription: Capturing Variability as a Source of Information." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7 (1): 8. http://dx.doi.org/10.5334/labphon.29.

Cooper, W., and P. Paccia-Cooper. 1980. *Syntax and Speech*. Cambridge, MA: Harvard University Press.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Delais-Roussarie, E., B. Post, M. Avanzi, C. Buthke, A. Di Cristo, I. Feldhausen, S.-A. Jun, et al. 2015. "Intonational Phonology of French: Developing a ToBI System for French." In *Intonation in Romance*, edited by Sonia Frota and Pilar Prieto, 63–100. Oxford: Oxford University Press.

Dilley, L. C. 2005. "The Phonetics and Phonology of Tonal Systems." PhD diss., MIT.

Dilley, L. C., M. Breen, M. Bolivar, J. Kraemer, and E. Gibson. 2006. "A Comparison of Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices)." In *Proceedings of Interspeech 2006*. https://www.isca-speech.org/archive/archive_papers/interspeech_2006/i06_1619.pdf.

Dilley, L., and M. Brown, M. 2005. "The RaP (Rhythm and Pitch) Labeling System." Version 1.0. https://pdfs.semanticscholar.org/5f73/1dbcafb2b64da6eb15daa67718866bc74cc9.pdf.

Dilley, L., S. Shattuck-Hufnagel, and M. Ostendorf. 1996. "Glottalization of Word-Initial Vowels as a Function of Prosodic Structure." *Journal of Phonetics* 24:423–444.

Dupoux, E., N. Sebastián-Gallés, E. Navarrete, and S. Peperkamp. 2008. "Persistent Stress 'Deafness': The Case of French Learners of Spanish." *Cognition* 106:682–706.

Escudero, D., L. Aguilar, M. del M. Vanrell, and P. Prieto. 2012. "Analysis of Inter-Transcriber Consistency in the Cat_ToBI Prosodic Labelling System." *Speech Communication* 54:566–582.

Face, T., and P. Prieto. 2007. "Rising Accents in Castilian Spanish: A Revision of Sp_ToBI." *Journal of Portuguese Linguistics* 6:117–146.

Fiscus, J., J. Garofolo, M. Przybocki, W. Fisher, and D. Pallett. 1998. *English Broadcast News Speech (HUB4)*. Philadelphia: Linguistic Data Consortium.

Fletcher, J., and J. Harrington. 1996. "Timing of Intonational Events in Australian English." In *Proceedings of the Sixth Australian International Conference on Speech Science and Technology*, edited by P. McCormack and A. Russell, 611–615. Australian Speech Science and Technology Association, Canberra, Australia.

Fodor, J. D. 1998. "Learning to Parse." *Journal of Psycholinguistic Research* 27:285–319.

Fodor, J. D. 2002. "Prosodic Disambiguation in Silent Reading." *Proceedings of the North East Linguistic Society* 32:113–132.

Frota, S. 2014. "The Intonational Phonology of European Portuguese." In Jun 2014b, 7–42.

Frota, S., P. Oliveira, M. Cruz, and M. Vigário. 2015. *P-ToBI: Tools for the Transcription of Portuguese Prosody*. Lisbon: Laboratório de Fonética, CLUL/FLUL. http://labfon.letras.ulisboa.pt/InAPoP/P-ToBI/.

Frota, S., and P. Prieto. 2015. *Intonation in Romance*. Oxford: Oxford University Press.

Garellek, M. 2013. "Production and Perception of Glottal Stops." PhD diss., UCLA.

Gee, J. P., and F. Grosjean. 1983. "Performance Structures: A Psycholinguistic and Linguistic Appraisal." *Cognitive Psychology* 15:411–458.

Godjevac, S. 2005. "Transcribing Serbo-Croatian Intonation." In Jun 2005c, 146–171.

Goldsmith, J. 1976. "Autosegmental Phonology." PhD diss., MIT.

Gordon, M. 2005. "Intonational Phonology of Chickasaw." In Jun 2005c, 301–330.

Grice, M., S. Baumann, and R. Benzmueller. 2005. "German Intonation in Autosegmental-Metrical Phonology." In Jun 2005c, 55–83.

Grice, M., and R. Benzmueller. 1995. "Transcription of German Intonation Using ToBI-Tones—The Saarbrücken System." *Phonus* 1:33–51.

Grice, M., M. D'Imperio, M. Savino, and C. Avesani. 2005. "Strategies for Intonation Labeling across Varieties of Italian." In Jun 2005c, 362–389.

Grice, M., and M. Savino. 2003. "Map Tasks in Italian: Asking Questions about Given, Accessible and New Information." *Catalan Journal of Linguistic* 2:153–180.

Gussenhoven, C. 2015. "Does Phonological Prominence Exist?" *Lingue e Linguaggio* 14:7–24.

Halliday, M. A. K. 1967. *Intonation and Grammar in British English*. The Hague, the Netherlands: Mouton.

Hayes, B. 1989. "The Prosodic Hierarchy in Meter." In *Perspectives on Meter*, edited by P. Kiparsky and G. Youmans, 203–260. New York: Academic Press.

Hirschberg, J., and J. Pierrehumbert. 1986. "Intonational Structuring of Discourse." *Proceedings of the 24th Meeting of the Association for Computational Linguistics*, 126–144.

Hirst, D. 1998. "Intonation in British English." In *Intonation Systems*, edited by D. Hirst and A. Di Cristo, 56–77. Cambridge: Cambridge University Press.

Hirst, D., and A. Di Cristo. 1998. "A Survey of Intonation Systems." In *Intonation Systems*, edited by D. Hirst and A. Di Cristo, 1–44. Cambridge: Cambridge University Press.

Hualde, J., and P. Prieto. 2016. "Towards an International Prosodic Alphabet (IPrA)." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7 (1): 1–25, especially 5. http://dx.doi.org/10.5334/labphon.11.

Indjieva, E. 2009. "Oirat Tones and Break Indices (O-ToBI). Intonational Structure of the Oirat Language." PhD diss., University of Hawaii.

Ito, K., and S. R. Speer. 2008. “Anticipatory Effect of Intonation: Eye Movements during Instructed Visual Search.” *Journal of Memory and Language* 58:541–573.

Jun, S.-A. 1998. “The Accentual Phrase in the Korean Prosodic Hierarchy.” *Phonology* 15 (2): 189–226.

Jun, S.-A. 2000. “K-ToBI (Korean ToBI) Labeling Conventions, Version 3.1.” *UCLA Working Papers in Phonetics* 99:149–173.

Jun, S.-A. 2005a. “Korean Intonational Phonology.” In Jun 2005c, 201–229.

Jun, S.-A. 2005b. “Prosodic Typology.” In Jun 2005c, 430–458.

Jun, S.-A., ed. 2005c. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, S.-A. 2010. “The Implicit Prosody Hypothesis and Overt Prosody in English.” *Language and Cognitive Processes* 25:1201–1233.

Jun, S.-A. 2014a. “Prosodic Typology: By Prominence Type, Word Prosody, and Macro-Rhythm.” In Jun 2014b, 520–539.

Jun, S.-A., ed. 2014b. *Prosodic Typology II: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Jun, S.-A., and J. Bishop. 2015. “Priming Implicit Prosody: Prosodic Boundaries and Individual Differences.” *Language and Speech* 58 (4): 459–473.

Jun, S.-A., and J. Fletcher. 2014. “Methodology of Studying Intonation: From Data Collection to Data Analysis.” In Jun 2014b, 493–519.

Jun, S.-A., and S. Kim, S. 2004. “Default Phrasing and Attachment Preferences in Korean.” In *Proceedings of Interspeech 2004*. https://www.isca-speech.org/archive/archive_papers/interspeech_2004/i04_3009.pdf.

Jun, S.-A., and C. Koike. 2008. “Default Prosody and RC Attachment in Japanese.” *Japanese-Korean Linguistics* 3:41–53.

Jun, S.-A., and M. Oh. 2000. “Acquisition of 2nd Language Intonation.” *Proceedings of International Conference on Spoken Language Processing* 4:76–79.

Karlsson, A. 2014 “The Intonational Phonology of Mongolian.” In Jun 2014b, 187–215.

Khan, S. D. 2008. “Intonational Phonology and Focus Prosody of Bengali.” PhD diss., UCLA.

Khan, S. D. 2014. “The Intonational Phonology of Bangladeshi Standard Bengali.” In Jun 2014b, 82–117.

Kjelgaard, M. M., and S. Speer. 1999. “Prosodic Facilitation and Interference in the Resolution of Temporary Syntactic Closure Ambiguity.” *Journal of Memory and Language* 40:153–194.

Klatt, D. H. 1975. “Vowel Lengthening Is Syntactically Determined in a Connected Discourse.” *Journal of Phonetics* 3:129–140.

Ladd, D. R. 1983. “Phonological Features of Intonational Peaks.” *Language* 59:721–759.

Ladd, D. R. 1990. “The Metrical Representation of Pitch Register.” In *Papers in Laboratory Phonology I*, edited by J. Kingston and M. Beckman, 35–57. Cambridge: Cambridge University Press.

Ladd, D. R. (1996) 2008a. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Ladd, D. R. 2008b. "Review of Prosodic Typology by Jun 2005c." *Phonology* 25 (2): 372–376.

Leben, W. 1973. "Suprasegmental Phonology." PhD diss., MIT.

Leben, W. 1976. "The Tones in English Intonation." *Linguistic Analysis* 2:69–107.

Lee, E.-K., and D. Watson. 2011. "Effects of Pitch Accents in Attachment Ambiguity Resolution." *Language and Cognitive Processes* 26 (2): 262–297. https://linguistics.ucla.edu/people/jun/papers%20in%20pdf/ASAPoster2016LeeandJun.pdf.

Lee, H., and S.-A. Jun. 2016. "Types of Errors in English Prosody Made by Native Korean Speakers." Poster presented at the Acoustical Society of America, Honolulu, Hawaii, December 1. https://linguistics.ucla.edu/people/jun/papers%20in%20pdf/ASAPoster2016YooandJun.pdf.

Lehiste, I. 1973. "Phonetic Disambiguation of Syntactic Ambiguity." *Glossa* 7 (2): 107–121.

Liberman, M. 1975. "The Intonational System of English." PhD diss., MIT.

Liberman, M., and J. Pierrehumbert. 1984. "Intonational Invariance under Changes in Pitch Range and Length." In *Language Sound Structure*, edited by M. Aronoff and R. Oehrle, 157–233. Cambridge, MA: MIT Press.

Lieberman, P. 1967. *Intonation, Perception and Language*. Cambridge, MA: MIT Press.

Maekawa, K., H. Kikuchi, Y. Igarashi, and J. Venditti. 2002. "X-JToBI: An Extended J-ToBI for Spontaneous Speech." In *Proceedings of the Seventh International Conference on Spoken Language Processing*. https://www.isca-speech.org/archive/archive_papers/icslp_2002/i02_1545.pdf.

Mayo, C., M. Aylett, and D. R. Ladd. 1997. "Prosodic Transcription of Glasgow English: An Evaluation Study of GlaToBI." In *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications*. https://www.cstr.ed.ac.uk/downloads/publications/1997/esca2.pdf.

McGory, J. 1997. "Acquisition of Intonational Prominence in English by Seoul Korean and Mandarin Chinese Speakers." PhD diss., Ohio State University.

Mennen, I. 1998. "Stability of Tonal Alignment: The Case of Greek Prenuclear Accents." *Journal of Phonetics* 26 (1): 3–25.

Mennen, I. 2004. "Bi-Directional Interference in the Intonation of Dutch Speakers of Greek." *Journal of Phonetics* 32:543–563.

Mennen, I. 2007. "Phonological and Phonetic Influences in Non-Native Intonation." In *Non-Native Prosody: Phonetic Descriptions and Teaching Practice*, edited by J. Trouvain and U. Gut, 53–76. Berlin: Mouton de Gruyter.

Mennen I. 2015. "Beyond Segments: Towards a L2 intonation Learning Theory." In *Prosody and Language in Contact*, edited by E. Delais-Roussaire, M. Avanzi, and S. Herment, 171–188. Heidelberg, Germany: Springer-Verlag.

Millar, J., J. Vonwiller, J. Harrington, and P. Dermody. 1994. "The Australian National Database of Spoken Language." *Proceedings of the ICASSP* 94:97–100.

Mo, Y., J. Cole, and E. K. Lee. 2008. "Prosody Perception by Naive Listeners: Evidence from a Large Multi-Transcriber Reliability Study." Poster presented at the Eighty-Second Annual Meeting of the Linguistic Society of America, Chicago, IL. January 3–6.

Nespor, M., and I. Vogel. 1986. *Prosodic Phonology*. Dordrecht, the Netherlands: Foris.

O'Brien, M., and U. Gut. 2010 "Phonological and Phonetic Realisation of Different Types of Focus in L2 Speech." In *Achievements and Perspectives in the Acquisition of Second Language*

Speech: New Sounds 2010, edited by K. Dziubalska-Kołaczyk, M. Wrembel, and M. Kul, 205–215. Frankfurt, Germany: Peter Lang.

Ostendorf, M., P. J. Price, and S. Shattuck-Hufnagel. 1995. *The Boston University Radio News Corpus*. Technical Report ECS-95–001. Boston: Boston University.

Ostendorf, M., I. Shafran, S. Shattuck-Hufnagel, L. Carmichael, and W. Byrne. 2001. "A Prosodically Labeled Database of Spontaneous Speech." In *Proceedings of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*. https://www.isca-speech.org/archive_open/archive_papers/prosody_2001/prsr_022.pdf.

Palmer, H. E. 1922. *English Intonation with Systematic Exercises*. Cambridge: Heffer.

Peng, S-H., C. Marjorie, C.-Y. Tseng, T. Huang, O. J. Lee, and M. Beckman. 2005. "Towards a Pan_Mandarin System for Prosodic Transcription." In Jun 2005c, 230–270.

Pierrehumbert, J. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pierrehumbert, J., and M. Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pierrehumbert, J., and J. Hirschberg. 1990. "The Meaning of Intonational Contours in the Interpretation of Discourse." In *Intensions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Pitt, M. A., K. Johnson, E. Hume, S. Kiesling, and W. Raymond. 2005. "The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability." *Speech Communication* 45:89–95.

Price, M. P. J., S. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. 1991. "The Use of Prosody in Syntactic Disambiguation." *Journal of Acoustical Society of America* 60:2956–2970.

Prieto, P. 2014. "The Intonational Phonology of Catalan." In Jun 2014b, 43–80.

Prieto, P., L. Aguilar, I. Mascaró, F. J. Torres-Tamarit, and M. Vanrell. M. 2009. "L'etiquetatge prosòdic Cat_ToBI.*" Estudios de Fonètica Experimental* 18:287–309.

Prieto, P., and M. Ortega-Llebaria. 2009. "Do Complex Tones Induce Syllable Lengthening in Catalan and Spanish?" In *Phonetics and Phonology: Interactions and Interrelations*, edited by M. Vigário, S. Frota, and M. J. Freitas, 51–70. Amsterdam: John Benjamins.

Prieto, P., and P. Roseano, eds. 2010. *Transcription of Intonation of the Spanish Language*. Munich: Lincom.

Schafer, A., J. Carter, C. Clifton, and L. Frazier 1996. "Focus in Relative Clause Construal." *Language and Cognitive Processes* 11:135–163.

Schafer, A., S. Speer, and P. Warren. 2005. "Prosodic Influences on the Production and Comprehension of Syntactic Ambiguity in a Game-Based Conversation Task." In *Approaches to Studying World-Situated Language Use*, edited by J. C. Trueswell and M. K. Tanenhaus, 209–225. Cambridge, MA: MIT Press.

Schafer, A., S. Speer, P. Warren, and S. White. 2000. "Intonational Disambiguation in Sentence Production and Comprehension." *Journal of Psycholinguistic Research* 29:169–182.

Schepman, A., R. Lickley, and D. R. Ladd. 2006. "Effects of Vowel Length and 'Right Context' on the Alignment of Dutch Nuclear Accents." *Journal of Phonetics* 34:1–28.

Scott, D. 1982. "Duration as a Cue to the Perception of a Phrase Boundary." *Journal of the Acoustical Society of America* 71:996–1007.

Selkirk, E. 1986. “On Derived Domains in Sentence Phonology.” In *Phonology Yearbook 3*, edited by C. Ewen and J. Anderson, 371–405. Cambridge: Cambridge University Press.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. “ToBI: A Standard for Labeling English Prosody.” In *Proceedings of the 1992 International Conference on Spoken Language Processing*. https://www.isca-speech.org/archive/archive_papers/icslp_1992/i92_0867.pdf.

Snedeker, J., and E. Casserly. 2010. “Is It All Relative? Effects of Prosodic Boundaries on the Comprehension and Production of Attachment Ambiguities.” *Language and Cognitive Processes* 25:1234–1264.

Speer, S., and A. Foltz. 2015. “Implicit Contrastive Prosody: Individual Differences in Processing and Production.” In *Explicit and Implicit Prosody in Sentence Processing: Studies in Honor of Janet Dean Fodor*, edited by L. Frazier and E. Gibson, 263–285. Berlin: Springer.

Speer, S. R., M. M. Kjelgaard, and K. M. Dobroth. 1996. “The Influence of Prosodic Structure on the Resolution of Temporary Syntactic Closure Ambiguities.” *Journal of Psycholinguistic Research* 25:249–271.

Stirling, L., J. Fletcher, I. Mushin, and R. Wales. 2001. “Representational Issues in Annotation: Using the Australian Map Task Corpus to Relate Prosody and Discourse Structure.” *Speech Communication* 33:113–134.

Thorsen, N. 1980. “A Study of the Perception of Sentence Intonation: Evidence from Danish.” *Journal of the Acoustical Society of America* 67:1014–1030.

Thorsen, N. 1983. “Two Issues in the Prosody of Standard Danish.” In *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd, 27–38. Heidelberg, Germany: Springer.

Trofimovich, P., and W. Baker. 2006. “Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech.” *Studies in Second Language Acquisition* 28 (1): 1–30.

Ueyama, M., and S.-A Jun. 1998. “Focus Realization in Japanese English and Korean English Intonation.” *Japanese-Korean Linguistics* 7:629–645.

Venditti, J. 1997. “Japanese ToBI Labeling Guidelines.” *Ohio State University Working Papers in Linguistics* 50:127–162.

Venditti, J. 2005. “J_ToBI Model of Japanese Intonation.” In Jun 2005c, 172–200.

Wightman, C., S. Shattuck-Hufnagel, P. Price, and M. Ostendorf. 1992. “Segmental Durations in the Vicinity of Prosodic Phrase Boundaries.” *Journal of Acoustical Society of America* 91:1707–1717.

Wong, W. Y. P., M. K. M. Chan, and M. E. Beckman. 2005. “An Autosegmental-Metrical Analysis and Prosodic Annotation Conventions for Cantonese.” In Jun 2005c, 271–300.

Yoo, H.-J., and S.-A. Jun. 2016. “Distribution of Accentual Phrase-Medial Tones in Seoul Korean.” Poster presented at the Acoustical Society of America, Honolulu, Hawaii December 2.

Commentary on Chapter 4: An Enhanced Autosegmental-Metrical Theory (AM$^+$) Facilitates Phonetically Transparent Prosodic Annotation

Laura C. Dilley and Mara Breen

Entia non sunt multiplicanda praeter necessitate.
(Translation from Latin: Entities should not be multiplied beyond necessity.)
—Occam's Razor, attributed to William of Ockham (ca. 1285–1349)

Introduction

We welcome this opportunity to respond to the well-organized, thoughtful essay by Jun in this chapter and to share our perspective on the ToBI* enterprise—where by ToBI*, we mean all tones and break indices (ToBI)-like annotation systems, including Mainstream American English MAE_ToBI, German GToBI, and so on—and how this enterprise fits in with the scientific study of tone and intonation in language. Much time has now passed—some forty years—since the core theoretic ideas behind ToBI* were put forward in groundbreaking, well-cited PhD dissertations at MIT by Goldsmith (1976), Pierrehumbert (1980), and Liberman (1975), which formed the core ideas in what has come to be known as autosegmental-metrical (AM) theory (Ladd 2008). Furthermore, more than twenty-five years have passed since the original MAE-ToBI was developed (Beckman and Hirschberg 1994; Beckman, Hirschberg, and Shattuck-Hufnagel 2005); this development included the third (and most recent) ToBI workshop in Columbus, Ohio, in 1993, which the first author of this commentary attended following her first undergraduate year at MIT. This long time span provides perspective on the strengths and weaknesses of ToBI*, as well as the theory that underlies it.

Our commentary aims to contextualize Jun's chapter by highlighting theoretical insights from forty years ago that led to the broad adoption of AM theory, an approach that has facilitated the discovery of important empirical insights about the cross-linguistic structure of intonation. We then show that several serious problems exist with traditional AM theory as it stands, leading to limitations on ToBI*'s value as a scientific tool. We argue that these problems can be clearly traced to a theoretical failure to prioritize consistent and transparent codification of the role of syntagmatic relationships in intonational phonology. Drawing on empirical evidence about the attested cognitive representations for pitch in the world's nonlinguistic communicative tonal systems (i.e., music), we propose a theoretical clarification of syntagmatic elements in intonational phonology, leading to a proposal for an enhanced AM theory, or AM$^+$. We show that attributing both syntagmatic and paradigmatic properties to tones provides a unifying account of multiple outstanding challenges in intonational phonological research that have not yet found a satisfactory explanation, including (i) the tonal composition of Greek prenuclear accents (Arvaniti, Ladd, and Mennen 1998), (ii) influences of contour shape and slope on perception of phonological contrasts (Barnes et al. 2012; D'Imperio 2000; Niebuhr 2007b), (iii) evidence against a nonmonotonic interpolation function account of F0 turning points on metrically nonprominent syllables (Dilley 2005; Dilley and Heffner 2013; Ladd and Schepman 2003), (iv) the lack of invariant timing in bitonal pitch accents (Dilley, Ladd, and Schepman 2005), (v) characterization of pointed versus plateau-shaped pitch accents (Niebuhr and Hoekstra 2015), and several others. Finally, we present the rhythm and pitch (RaP) prosodic transcription system as an AM$^+$-based empirical tool that can be extended toward the goal of developing an international prosodic alphabet (IPrA; Hualde and Prieto 2016).

Enduring Insights from over Forty Years of Traditional AM Theory

Elaborating on Jun, we highlight some key ideas and findings from the last forty or more years that constitute contributions of ToBI* to knowledge about tonal aspects of linguistic systems:

- *Tones are autonomous from segments*: That tones are autonomous from segmental structures but temporally coordinated with them was a foundational idea for the field of intonational phonology, as highlighted by Jun. This idea provided the basis for ToBI*'s descriptive notations, in which entities such as H (high) and L (low) are viewed as discrete tonal events that have abstract associations with segments (e.g., Ladd 2008).
- *Surface intonation contours reflect sparse tonal representations*: Another core idea highlighted by Jun is the idea that tones are sparse; for example, they do not occur on every syllable and are connected via F0 interpolations.[1]
- *Prominence- and boundary-related tones have distinctive distributional properties*: The key idea of ToBI* that tones participate in either pitch accents or edge tones has stood the test of time. As highlighted by Jun, starred tones of pitch accents associate with (and unstarred tones flank) metrically prominent positions, while phrase tones associate with constituent edges.
- *Peaks, valleys, and elbows are phonologically significant evidence of tones*: Abundant evidence that falls largely outside the scope of Jun's essay has shown that in general, F0 peaks, valleys, and elbows—transitions from a flat region of pitch to a rise or fall—constitute phonologically significant evidence of "tones" across a wide variety of intonation languages (D'Imperio, Gili Fivela, and Niebuhr 2010; del Giudice et al. 2007; Knight and Nolan 2006; Welby 2006). These F0 points have been argued to serve as "control points" in production and to be important for perception (Gussenhoven 2004; House 1990; Ladd 2008). Furthermore, considerable evidence suggests that effects of abstract tonal structure on F0 are better conceived in terms of perceptual targets involving auditory pitch (Barnes et al. 2012; D'Imperio 2000). Jun hints at some problems with the ToBI* framework's handling of accounting for F0 turning points and facts about the importance of pitch for phonology, a topic we explore here.
- *Tones have paradigmatic phonological status*: A core proposal of both Goldsmith (1976) and Pierrehumbert (1980) was that tones have paradigmatic phonological status, meaning that they are defined relative to the speaker's pitch range. This proposal is supported by the observation that in lexical tone languages, perceivers can recognize the tone of a single-syllable word spoken in isolation with a level tone (Lee 2009; Peng et al. 2012). Relatedly, perceptual studies demonstrate that in intonation languages, listeners can discern the location of a syllable in a speaker's pitch range with reasonably good accuracy (Bishop and Keating 2012; Honorof and Whalen 2005).

- *Starred tones of pitch accents form associations with syllables that have hierarchical metrical prominence*: The theory behind ToBI* posited a notion of *starred tones*, that is, tones that participate in pitch accentuation by associating with metrically prominent syllables. The potential influence of the hierarchical organization of stress on tones that was first worked out in Liberman (1975) was not explored in Pierrehumbert (1980). However, the explanatory value of viewing stress as hierarchical and metrical survives to the present day.

We agree with Jun that these theoretical points capture important generalizations about tonal systems made possible by the invention of ToBI*. However, in the next section, we argue that Goldsmith's (1976) and, later, Pierrehumbert's (1980) assumption that tones have (strictly) paradigmatic phonological status provided an incomplete assessment of phonological properties of tone. We identify this theoretical choice as the source of considerable, enduring problems with ToBI*'s phonetic transparency and consistency.

Strictly Paradigmatic Phonological Representations Lead to Descriptive Inadequacy and Inconsistency in Traditional AM Theory and ToBI*

Jun alludes to theoretical problems with ToBI* by stating, "Some of [the] challenges [of ToBI*'s handling of intonational phenomena] stem from properties of the AM theory that ToBI adopts" (section 4.5). Jun cites, without elaboration, the lack of phonetic transparency in ToBI* to be one of its key problems. In this section, we trace ToBI*'s problems with phonetic transparency and consistency to inadequate treatment of syntagmatic aspects of tonal phonological representations.

It is abundantly clear that both paradigmatic aspects as well as syntagmatic aspects of representations are important for tonal systems (Cutler, Dahan, and van Donselaar 1997; Ladd 2008; Lee 2009). *Syntagmatic properties*, which involve defining tone height in relation to adjacent tones rather than to a global referent, have long been thought to be central to tonal representations across languages (Cole 2015; Jakobson, Fant, and Halle 1952; Ladd 2008; O'Connor and Arnold 1973; Odden 1995). There is considerable evidence that cognitive representations of tonal information include syntagmatic relationships in lexical tone languages (Odden 1995; Wong and Diehl 2003), intonation languages (Dilley 2005; Dilley and Brown 2007), and nonlinguistic tonal systems, such as world musical traditions (Burns 1999; Dowling and Fujitani 1971; Monelle 2014; Patel 2010).

Both Goldsmith (1976) and Pierrehumbert (1980) acknowledged the importance of syntagmatic relationships for tonal representations, but they prioritized only the capture of paradigmatic aspects in phonology. We will show that the assumption of strictly paradigmatic features in phonology was highly problematic. Still, given that Goldsmith (1976) marked the birth of the idea of true tonal autonomy from segments, it was arguably not the time to explore the specific featural representations of tones themselves.[2] Indeed, no linguistic theoretic notational device had yet been developed that could yield conceptual insight into how tones themselves can be dually paradigmatic *and* syntagmatic. (AM$^+$ develops such a device; see the "A Way Forward" section.)

The choice of strictly paradigmatic tonal phonological representations was viewed as a simplifying assumption, but this assumption led to an overall theory of the "grammar" that was, in practice, not simple. To justify delving into the sequelae of this theoretic choice, especially the explanatory burden put on the "phonetic component" of the grammar by assuming a very weak phonology, we cite Pierrehumbert and Beckman (1988, 4), who state, "the division of labor between the phonology and the phonetics is an empirical question, one which can only be decided by constructing complete models in which the role of both in describing the sound structure is made explicit." As we will discuss, the theoretic assumption that phonology lacks syntagmatic phonological restrictions on tones led to the following: (i) complex phonetic rules for tone scaling, which did not, in the end, "work" to achieve desired restrictions on relative tone heights; (ii) inconsistencies in assumed mappings of pitch accentual tones to significant F0 events (peaks and valleys); and (iii) complications in when F0 events corresponded to interpolation functions versus phonological tones (accents or phrase tones). These problems have led to difficulties in using ToBI* for prosodic typology (Hualde and Prieto 2016).

Complex Phonetic Rules and Mechanisms for Tone Scaling That Didn't Work

Accepting as true the a priori premise that tonal representations lack syntagmatic restrictions required complete redefinitions of what constitutes a "phonological representation" and what is "phonetic." That is, given the a priori premise that the phonological component of the grammar encodes only paradigmatic aspects of tones, the logical consequence was the further assumption that the phonetic component of the grammar is home to syntagmatic restrictions on relative tone heights. There is abundant evidence that syntagmatic changes—being higher or lower than another tone—are meaningful, and until Pierrehumbert (1980), meaningful contrasts were considered to be part of phonology. Suddenly, the phonetic component of the grammar, which prior to that time had been taken to refer to, for example, biomechanical forces during speech production, was endowed with the power to make meaning-based distinctions.[3]

To supplement this "weak" phonology, it was necessary to invent a "strong" phonetics that consisted, in Jun's words, of "rules that map the phonological representation (abstract level tone target sequences) to the phonetic representation (the F0 contour)" (section 4.2). These rules, comprising a complex set of equations laid out in an entire chapter of Pierrehumbert (i.e., Chapter 4), were the main mechanism in the "grammar" for scaling the relative F0 heights of tones, one to another. They entailed an assumption of an abstract tone reference line necessary for phonetic scaling of tones, together with a gradient parametric value (which was termed "prominence" but was equated with F0), along with abstruse parameters *n* and *k*, which lacked a phonetic interpretation. Pierrehumbert and Beckman (1988) later proposed a version of the phonetic module that dispensed entirely with the phonetic rules, instead proposing that paradigmatic tones were scaled with respect to both a high reference line and low reference line, as a function of a parameter again termed "prominence" but which was just a proxy for F0. A variety of other proposals were put forward that varied with respect to numbers of reference lines, whether reference lines were static or dynamically changed, and whether tones were assumed to be on reference lines or could vary freely with respect to the reference lines (e.g., Ladd 1986). In many cases, the reference lines were just a proxy mechanism for imposing syntagmatic (phonological) restrictions on relative tone heights, as in Liberman and Pierrehumbert (1984). These accounts ignored the issue of how listeners could perceptually recover phonological representations from F0, or else they sidestepped the issue by assigning meaning to the "phonetic" component rather than to phonology.

There was, furthermore, a serious problem with the phonetic rules in Pierrehumbert (1980): they did not actually restrict syntagmatic relative F0 heights of tones. As demonstrated in Dilley and Brown (2007, 545–548), the rules failed to successfully restrict scaling of L and H tones so that specific claimed F0 contours would correspond to the intended tonal entities. For example, Dilley and Brown showed that even for bitonal accents such as L + H* and L* + H (uniformly assumed to entail rising contours), the rules permitted H tones to fall below the adjacent L tones, allowing L + H* and L* + H to map onto falling contours. The revised theory of Pierrehumbert and Beckman (1988) also suffered from the same serious problem, as further shown in Dilley and Brown (548–549), so that again, rising portions of L + H* and L* + H contours were permitted to map onto falling contours. Dilley and Brown showed that problems of this sort are not limited to these two accents, but are instead widespread throughout the accounts for tonal sequences of a variety of types.

Inconsistencies in Mapping Pitch Accents to F0 Events

Numerous complications and inconsistencies in the pitch accent inventory can be traced to the piecemeal way in which syntagmatic restrictions were handled in Pierrehumbert (1980). The theoretical distinction between bitonal accents such as L + H* and single-tone accents such as H* was itself motivated in part as a means of capturing syntagmatic relations. Specifically, Pierrehumbert states, “A pitch accent *can impose a particular relationship between the f0 on the accented syllable and the immediately preceding or following f0 value*, independent of the existence of other accents. . . . In our theory, *the bitonal accents* [H* + L, H + L*, L* + H, L + H*] *have this property* and there are also two single tones [H*, L*] which do not” (31, our emphasis). Note that this treatment implicitly posited that relative heights of other tones in sequence (for example, L* followed by H*) were unconstrained in their relative heights by phonology, leaving a legacy of inconsistent treatment in ToBI*’s notational conventions regarding which pairs of tones in a sequence code for syntagmatic relative tone heights, and which do not. As we already noted, the tone scaling rules did not actually restrict the syntagmatic relative heights of the two tones of bitonal pitch accents to surface with the intended F0 contours.

The piecemeal handling of syntagmatic restrictions further complicated the treatment of pitch accents through the adoption of descriptive devices termed “floating low” tones. For example, the H* + L bitonal accent in Pierrehumbert was treated as exceptional, in that the +L tone was assumed to be “floating.” It was assumed to be never directly realized phonetically as a low F0 event, but instead was assumed to be the causal factor in an observed F0 peak (which is normally thought of as an index of an H tone) being relatively lower than another F0 peak in the same phrasal constituent.[4] This indirect floating-low device as a means of accounting for a syntagmatic relationship among observed high-pitched events was borrowed from mid-1970s African linguistics (and Goldsmith 1976), according to which lexical L tones were sometimes associated with and/or synchronically traceable to observed lowering of subsequent H-toned units, resulting in iterative phonetic lowering of the H-tone syllables in a claimed phenomenon termed “downstep.”

Complications in When F0 Curves Correspond to Phonetic Interpolation versus Tones

The exceptional treatment of the L tone in H* + L accents as a floating-low tone in Pierrehumbert (1980) in order to codify a syntagmatic relationship among high-toned events necessitated a further theory-internal complication regarding F0 interpolation contours. Building on a core assumption that phonological specification of tones is usually sparse in intonation languages, Pierrehumbert proposed that, in general, F0 interpolation functions that connect phonological tones are monotonic: increasing functions should only increase, not decrease, and decreasing functions should only decrease, not increase. Because it was assumed that the L in an HLH sequence was a floating low that could never "surface" as an F0 valley, this precluded any description that treated the F0 valley as a low tone when the following peak was not lower than an earlier peak. The theoretic choice to prioritize the descriptive device of floating low from mid-1970s African linguistics over phonetic consistency meant that for an F0 peak-valley-peak sequence in which the two peaks were of equal height, the F0 valley could not be described as an L tone between the two H tones, based on these theory-internal assumptions. It was therefore necessary for Pierrehumbert (1980) to posit an exceptional nonmonotonic ("sagging") interpolation function only in the case of two H tones, when the second H tone was not lower.

Evidence against this function was demonstrated in Ladd and Schepman (2003). Specifically, not only did the F0 valley in question show consistent alignment with respect to the phonological structures in utterances, but varying the temporal alignment of the low tone within a phrase changed listeners' interpretation of that phrase. Both kinds of evidence were consistent with the F0 valley being a reflex of a low tone that was phonological in nature. To further complicate matters, Pierrehumbert (1980) assumed H tones sometimes were realized with a "late peak" on a nonprominent syllable following the accented syllable. This assumption constitutes a lesser-known exception to the monotonic interpolation rule, one not commented on by Pierrehumbert, and amounts to a second type of nonmonotonic function, termed a "bulging interpolation" by Dilley and Heffner (2013).

Moreover, examination of how phonological theories have handled cases of phonetically flat pitch reveals another case of how failing to codify syntagmatic relationships has complicated theories. Consider that *monotonic* can also mean "unchanging in pitch or tone"; a monotonic interpolation between two tones at the same level should yield a flat pitch, where a temporally later tone has an equal pitch relative to an earlier tone and to everything in between.[5] To account for regions of flat pitch, descriptive work on African languages in the 1970s (e.g., Goldsmith 1976; Hyman and Schuh 1974; Leben 1973; Williams 1971/1976) posited phonological rules that enacted tone copying or spread to account for regions of flat pitch, that is, cases where lexically specified H or L tone showed a sustained pitch at the same level over multiple syllables.[6] We note that tone spread is assumed to result in an F0 "elbow" that marks the right edge of the flat-pitched region before a subsequent rise or fall.[7]

An alternative to the tone-spreading rule proposed by Pierrehumbert and Beckman (1988) involved accounting for syntagmatically level stretches of F0 according to a phonological rule known as "secondary association," in which a single tone could be anchored to two timing slots separately. This idea was productively used to account for level stretches in a variety of languages (Grice 1995; Grice, Ladd, and Arvaniti 2000; Prieto, D'Imperio, and Gili Fivela 2005). However, note that this proposal (and tone spreading) requires inconsistency in treatment of autosegmental association. That is, while the original idea of Goldsmith (1976) was that tones occupy a single timing slot (i.e., that they occur at a single moment in time), secondary association entails that tones can occupy multiple timing slots and "persist" over long stretches of time.

The tension among the tone spreading account, the secondary association account, tone copying, and/or a single tone per timing slot with monotonic interpolation has not to date been resolved. The core facts motivating these proposals, however, were strikingly similar. That is, cross-linguistically, there are many attested cases in which an equal height relationship exists among successive, adjacent tones, where change points may be separated by long distances.

Summary

In conclusion, the assumption of strictly paradigmatic tonal phonological representations had a cascade of negative consequences for phonetic transparency and theoretic consistency. We point to the obfuscation of syntagmatic relationships as an underappreciated, but truly fundamental flaw in traditional AM theory and the ToBI* enterprise. A span of forty years' time also reveals that a strictly paradigmatic phonological treatment is simply inadequate. In spite of the best efforts of Pierrehumbert (1980), Pierrehumbert and Beckman (1988), and others, a supplementary phonetic module has not been put forward that sufficiently constrains relative tone heights to generate the correct F0 curves from phonological tones. The legacy of this inadvertent obfuscation masquerading as theoretical simplification has indelibly imprinted in ToBI*'s notational apparatus. Failure to clearly codify the relationship between syntagmatic aspects in the signal and abstract theoretical constructs means that ToBI's descriptive apparatus for these aspects of representations is highly inconsistent, unconstrained, and unprincipled.

Fortunately, ToBI* systems have been used by communities of scholars as though syntagmatic tonal relationships are part of the phonology, even though they are not. For example, scholars have annotated L+H* as a low valley plus a rising pitch, even though this choice is not supported by the underlying theories, as discussed. In the following section, we demonstrate that a simple theoretical change—to assume that syntagmatic features are directly part of the phonological representations of tones—allows building on the last forty years of insights in an "enhanced" AM framework.

A Way Forward: AM$^+$ Theory and the RaP Transcription System

An "enhanced" AM theory (AM$^+$) is proposed here. AM$^+$ integrates insights from more than forty years of empirical work in intonational phonology, as well as research in speech perception, music cognition, and cognitive neuroscience, building on the proposals of Dilley (2005). These proposals develop a notational device adapted for linguistic systems that is derived from insights about cognitive representations of nonlinguistic tonal information from auditory streaming studies, music cognition studies, and music theories for the world's musical systems (Bregman 1994; Burns 1999; Dowling and Fujitani 1971; Hannon and Trainor 2007; Jones, Fay and Popper 2010; Patel 2010).

A central part of the AM$^+$ theory is its assumption that, across languages, cognitive representations of tonal systems include both syntagmatic and paradigmatic aspects of tone. That is, they are both part of the phonology. Furthermore, paradigmatic and syntagmatic aspects of tone operate according to similar constraints such that they are specified lexically in some tonal systems and postlexically in others. It is proposed that each language draws on a combination of paradigmatic and syntagmatic tonal specifications, where there will be different densities of specification at the lexical or postlexical levels.[8]

Note that by including both paradigmatic and syntagmatic aspects of tone in phonology, AM$^+$ theory is not more complex than proposals of traditional AM theory, which assumed strictly paradigmatic tonal features (e.g., Goldsmith 1976). This is because, as we show for AM$^+$ theory, paradigmatic aspects of tone reduce to syntagmatic feature specifications. This is a core insight of AM$^+$ theory—namely, that paradigmatic tonal features can be formally reexpressed as syntagmatic ones. AM$^+$ thus preserves the economy of featural specification that was appealing in traditional AM theory. Furthermore, in nearly forty years, no theory based on paradigmatic level tones plus phonetic implementation rules has been put forward that successfully maps sequences of H and

L tones to their expected F0 outputs for intonation languages, as was conclusively shown in Dilley and Brown (2007).[9]

AM^+ conceives of tones in cognitive, abstract terms. In this theory, tones are abstract pitch targets that involve language-specific sensorimotor mappings. Conceiving of tones as abstract pitch targets that instantiate experience-dependent sensorimotor mappings is well grounded in empirical research from the past two decades in speech perception, music cognition, and cognitive neuroscience (Burnett et al. 1998; Chen et al. 2007; Guenther 2016; Guenther and Hickok 2015; Hutchins and Peretz 2011; Ning, Shih, and Loucks 2014; Patel et al. 2011; Pfordresher et al. 2015). To relate concepts of tones in traditional AM theory to AM^+, note that an H tone, which in traditional AM theory was taken to correspond to an F0 peak (see Pierrehumbert 1980), can be fundamentally reexpressed as an abstract pitch target that is syntagmatically constrained to be higher in pitch than a tonal target to the left and to the right. Viewed in this way, a syllable that is autosegmentally associated with an H tone naturally maps in most speaking situations to an F0 peak. However, because tonal targets are intrinsically perceptual in nature, other F0 mappings are possible, such as F0 plateaus (Dilley and Brown 2007; Knight 2008) or variations in the F0 shape as given by, for example, the tonal center of gravity (Barnes et al. 2012; D’Imperio 2000; Niebuhr 2007a).

Likewise, an L tone that in traditional AM theory was taken to correspond to an F0 valley (e.g., Pierrehumbert 1980) can be reexpressed in AM^+ as an abstract pitch target that is constrained to be syntagmatically lower in pitch than a tonal target to the left and to the right. Finally, an L or H tone that in traditional AM theory was taken, for example, to correspond to the right edge of a stretch of level pitch, is reexpressed in AM^+ as an abstract pitch target that is constrained to be syntagmatically at the same pitch level as a tonal target to the left. The syntagmatic relationship between that L or H tone and the tone that follows will then dictate whether the contour subsequently rises or falls (i.e., a following tonal target that is higher or lower, respectively). Given that perceptual pitch lawfully relates to F0 in speech (d’Alessandro and Mertens 1995; House 1990; Mertens 2004), AM^+ provides a unifying explanation for observed correspondences between abstract tones and their typical F0 consequences, as with F0 peaks, valleys, and plateaus.[10]

An experiment from Dilley and Brown (2007) provides further support for the proposal that categorical differences in F0 turning point timing, for example, of F0 peaks and valleys, derive fundamentally from pitch targets whose cognitive representations involve syntagmatic specifications. Dilley and Brown created synthetic stimuli with flat, level-pitched F0 across critical syllables, without F0 peaks or valleys. Using an imitation task, the gold standard test of categories in intonation (Gussenhoven 2004; Pierrehumbert and Steele 1989), Dilley and Brown showed that speakers imitated the level-pitched syllables by producing categorical shifts in F0 peak and valley timing; furthermore, the categorical timing was predicted by the syntagmatic relationship of relative height borne by a level-pitched syllable to adjacent syllables, not by the syllable’s relation to the pitch range. These findings further supported a view that F0 peaks and valleys derive from syntagmatic relationships among tones and provide experimental evidence for a tonal phonology that includes syntagmatic features.

The RaP prosodic transcription system (Breen et al. 2012; Dilley and Brown 2005) instantiates the proposals of AM^+ theory.[11] The phonological representations in AM^+ are based on two syntagmatic tone features: [+/– same], which distinguishes same and different, and [+/– higher], which distinguishes higher and lower. [+/– higher] is only specified in the case of [–same]. RaP includes the symbols **H**, **L**, and **E**, which capture the syntagmatic relationship borne by a tone, T_n, with respect to a previous tone, T_{n-1};

boldface type will be used for RaP symbols to distinguish them from ToBI* notations (in this section, for MAE_ToBI, in particular).[12] RaP's **H** designates a tone that has a feature specification [–same, +higher] and is phonetically higher than the previous tone. **L** designates a tone that has a feature specification [–same, –higher] and is phonetically lower than the previous tone. **E** designates a tone with a feature specification [+same], which is phonetically equal in pitch to the previous tone.[13]

In RaP, the features [+/– same] and [+/– higher] are specified for pairs of adjacent tones, T_{n-1} and T_n, on an AM$^+$ grid tier. In other words, it is claimed that the cognitive representation for tone in languages entails phonological encoding of the relative heights of tones in a sequence. The most basic aspect of the representation is that a given tone, T_n, is specified to have a pitch value that is higher than, lower than, or equal to that of a prior tone in the sequence, T_{n-1}. An AM$^+$ grid tier is a hybrid concept that generalizes across notions of a metrical grid row (e.g., Halle and Idsardi 1995) and an autosegmental tier à la Goldsmith (1976); it conceives of autosegmental association as expressly hierarchical, elaborating on the assumed, and eponymous, metrical representations of traditional AM theory. Furthermore, the notation T_n/T_{n-1} is adopted to represent a pair of adjacent tones on an AM$^+$ grid tier that is constrained by a given syntagmatic feature; the entity on the right of the / is the referent entity. For example, T_n/T_{n-1} = [–same, +higher] means that T_n is higher than T_{n-1}. Phonetically, this corresponds to a rise. By extension, a reciprocal relationship exists between two tones captured through the relationality of this expression. A rise in forward time is just a fall in reverse time, which is captured by a sign change when the referent entity is in the future, for example, T_{n-1}/T_n = [–same, –higher]. In AM$^+$ theory, this is termed the *reciprocal property*.[14]

Paradigmatic features have been traditionally characterized as "tone levels" according to which tones are defined relative to a speaker's pitch range. AM$^+$ offers a formalization of this view according to which paradigmatic tone levels arise from a syntagmatic relationship between a tone, on the one hand, and an abstract (phonological) referent quantity, on the other, which is phonetically defined with respect to a speaker's own pitch range. Specifically, paradigmatic tonal representations are formally codified as a syntagmatic relationship between a lexically specified tone, T, and an abstract referent level, *r*; the value *r* is phonetically interpreted as the speaker's mean pitch (or habitual pitch).[15] A high tone which is high in speakers' pitch ranges is represented as T/r = [–same, +higher], a low tone which is low in speakers' ranges is T/r = [–same, –higher], and a tone at speakers' mean or habitual pitch levels is T/r = [+same].[16] If a tone, T, is not specified in the lexicon to have a particular featural relationship with respect to *r*, then at the speech motor planning stage, we propose that the first tone in an utterance, T_1, receives postlexical assignment of features for T_1/r. Thereafter, lexically specified features for tones, together with postlexical expressive factors such as prominence and intended meaning, will determine the overall placement of tones in the speaker's pitch range and the syntagmatic pitch distances among tone pairs.

Importantly, paradigmatic representations specified according to a common referent have an interesting benefit: they allow obtaining syntagmatic relationships "for free" when tones are strung together by default in sequence.[17] For example, a language with two lexical tones, T_H for high tone and T_L for low tone, might specify that T_H/r = [–same, +higher] and T_L/r = [+same].[18] Because T_H is higher than *r* and T_L is at the same level as *r*, deductive reasoning ensures that by default, T_H will be higher than T_L. Language-specific rules might modify default syntagmatic relationships in ways that could be used to distinguish meanings (Odden 1995). This account appears to fit well in the case of Hausa, where syntagmatic relative heights of H tones in HL sequences distinguish statements from questions (Inkelas and Leben 1990; Inkelas, Leben, and Cobler 1986).

These proposals entail that syntagmatic relationships can be derived from paradigmatic specifications, or syntagmatic relationships can even be separately specified in the lexicon or the morphosyntax, perhaps without a paradigmatic specification. These properties therefore endow this framework with the ability to more elegantly account for so-called floating-tone phenomena than was previously possible. Specifically, if a tone-bearing unit of a given tone is deleted, there is still another tone that retains, or holds onto, the syntagmatic relational featural specification. A new anchor point for this abstract relational structure can then be found, leading to realization on a different syllable the relative height differences taken to be phonetic hallmarks of floating tones, which are here viewed as reflexes of inherently syntagmatic tonal specifications.

Elaborating on Dilley (2005), five tonal levels can be captured in AM^{+} by proposing the feature [+/– small]. This feature codifies tonal distance: [+small] represents a small tonal distance, while [-small] indicates a large tonal distance (Patel 2010; Vos and Troost 1989). We propose that [+/–small], like [+/–high], is specified only for [–same]. We further propose that in most cases, tones are unmarked for [+/–small], so pitch range can vary expressively. A language with five level tones—extra high, high, mid, low, extra low (EH, H, M, L, EL, respectively)—could thus be described as in table 4c.1.[19]

RaP and AM^{+} theory elaborate productively on the relationship between hierarchical metrical structure and tonal associations. AM^{+} and RaP adopt the starred * tone notation used previously to describe tones that autosegmentally associate with a metrically prominent syllable. Metrically prominent syllables are marked in RaP with **x** (moderate prominence) or **X** (strong prominence), where the latter would occupy a higher grid tier position than the former. Importantly, AM^{+} theory proposes that starred tones that associate with prominent metrical positions propagate upward to be represented in positions of adjacency on higher grid tiers. Following the idea of traditional metrical grid formalisms (e.g., Halle and Idsardi 1995; Hayes 1995), higher levels of AM^{+} grid tiers entail adjacency of elements that occupy them. The significance of this is that on higher grid tiers, nonadjacent tones may be specified for syntagmatic featural relationships lexically or postlexically. This allows an account of tone register phenomena, for example, downstep, downdrift, and upstep (Clements and Goldsmith 1984; Hyman 1993; Inkelas and Leben 1990; Inkelas et al. 1986; Ladd 1988; Snider 1999; Truckenbrodt 2002).[20] A metrical account is consistent with a growing body of

Table 4c.1
Five level paradigmatic tone specifications derived from syntagmatic features [+/– same], [+/– higher], and [+/– small]

Lexical specification in phonology	Phonetic interpretation
T_{EH}/r = [–same, +higher, –small]	Substantially higher than the mean pitch; high in the pitch range
T_{H}/r = [–same, +higher, +small]	Slightly higher than the mean pitch
T_{M}/r = [+same]	Equal to the mean pitch
T_{L}/r = [–same, –higher, +small]	Slightly lower than the mean pitch
T_{EL}/r = [–same, –higher, –small]	Substantially lower than the mean pitch; low in the pitch range

evidence of metrical interactions in a variety of languages with very different tonal systems (de Lacy 2002; Hayes 1995; Manfredi 1993; Rice 1987; Zec 1999).

Phonetically, RaP requires a local phonetic pitch change (F0 turning point or F0 slope change) for a starred tone to be indicated on a metrically prominent syllable. This is consistent with prior assumptions of sparse representations, e.g., that only a subset of metrically prominent syllables are pitch accented (i.e., associated with a starred tone). As a consequence, a stretch of flat pitch can never have pitch accents, only metrical prominences, in contrast to Pierrehumbert's (1980) proposal that strings of low pitch accents can be present in regions of flat pitch. Based on these ideas, RaP distinguishes three categories of syllable prominence: metrically nonprominent, metrically prominent without a pitch accent (i.e., without a pitch change, such as for flat pitch), and metrically prominent with a pitch accent (i.e., with a pitch change). By contrast, the ToBI* system allows for only two levels of prominence, pitch accent or no pitch accent, in contrast to multiple studies demonstrating that speakers produce, and listeners perceive, at least three levels of prominence (Fitzroy and Breen 2019; Greenberg, Carvey, and Hitchcock 2002).

There are several other notational conventions and standardizations that are instantiated in AM$^+$ and codified in RaP's conventions which enhance phonetic transparency and explanatory power relative to ToBI*:

- *Strictly monotonic interpolation functions*: AM$^+$ retains the assumption that tones may be sparsely distributed in intonation languages and do not often occur on every syllable or word, such that adjacent tones separated by multiple syllables or words may be connected via interpolations. In this theory, interpolation functions are strictly monotonic, ensuring that all turning points are coded as tones. Multiple studies have demonstrated evidence against Pierrehumbert's (1980) proposal that certain F0 turning points are not tones but rather reflexes of exceptional nonmonotonic interpolation functions. (See Dilley and Heffner 2013; Dilley, Ladd, and Schepman 2005; Ladd and Schepman 2003.)
- *Tones and timing slots*: An aspect of AM$^+$ theory that notably increases phonetic transparency is that every tone must be associated with a timing slot. This effectively disallows floating tones and multiple associations between a single tone and more than one timing slot (tone spread or secondary association). Like ToBI*, RaP allows multiple tones to be associated with a syllable.

The assumption that every tone is associated with a timing slot allows for the consistent treatment of unstarred tones in bitonal pitch accents. Pierrehumbert (1980) predicted a constant timing relationship between the two tones of bitonal pitch accents, but this prediction has not been borne out in production studies (Arvaniti, Ladd, and Mennen 1998, 2000; Dilley, Ladd, and Schepman 2005; Ladd 2008). Following AM$^+$, RaP treats pitch accents as prominence-lending pitch movements, that is, locations of a local change in pitch. RaP assumes that unstarred tones can participate in pitch accents by associating with a nonprominent slot adjacent to a timing slot with a starred tone, or they can associate with constituent edges, following Pierrehumbert and Beckman (1988). The + symbol is used in RaP for the unstarred tones of pitch accents. Unstarred tones with + are assumed to associate directly with respect to a metrically nonprominent position; however, their eligibility to so associate is limited to the set of nonprominent positions that are adjacent to a prominent position associated with a starred tone. The + is put on the right side of the unstarred tone when that tone is to the left of a starred prominent syllable (i.e., a metrically prominent syllable that is autosegmentally associated with a starred tone). For example, RaP's **L+ H*** is annotated as a sequence of two tones with a space between them and maps to ToBI*'s L + H*. Otherwise, the + is put on the left side of the tone, as in RaP's **L* +H**, which can be compared to ToBI*'s L* + H. Note that this formulation predicts that unstarred tones can be associated with metrically nonprominent positions both before and after a starred tone.

- *Meaningful pitch range differences*: The theory outlined here, which places primacy on syntagmatic tone features, readily accounts for examples of meaningful differences in pitch range. For example, RaP accounts for ToBI's much-studied distinction of H* versus L + H*; phonetically, there is a small rise to a peak for ToBI's H* versus a large rise to a peak for ToBI's L + H*.[21] These phonetic differences are important for capturing distinctions of focus (Breen et al. 2010; Féry and Krifka 2008; Katz and Selkirk 2011; Xu and Xu 2005). Noting that both contours rise to a peak, RaP captures the contours as **L+ !H*** (for ToBI's H*) versus **L+ H*** (for ToBI's L + H*). This treatment has the further effect of filling a theoretical gap in Pierrehumbert's (1980) theory having to do with the status of F0 values on phrase-initial unstressed syllables preceding an H*. When H* was on a noninitial syllable in the phrase and there was no phrase-initial boundary tone (such as Pierrehumbert's %H), it was theoretically unclear how phrase-initial unstressed syllables could obtain an F0 value under the theory, because there was no phrase-initial tone prior to the H* with respect to which to carry out F0 interpolation (Dilley 2010). RaP assumes that every phrase starts and ends with a tone, thus overcoming this theoretical gap.
- *Phrase edges*: Regarding phrase-initial tones, a further point is warranted about RaP notation. Recall that in RaP, the codes **H**, **L**, and **E** describe the syntagmatic relationship between the later-occurring tone and the earlier-occurring one. By definition, a phrase-initial tone has no earlier-occurring tone in the same phrase; rather, its phonological status as high or low is fully determined by the following tone (if there is no paradigmatic lexical specification, that is). The phrase-initial tone thus redundantly corresponds to the reciprocal (via the reciprocal property) of the tone in second position in the phrase. This status of the phrase-initial tone is designated by prepending the : symbol. That is to say, the three ways of beginning a phrase are a rise, symbolized **:L H** (omitting + and *); a fall, symbolized **:H L**; or a level pitch, symbolized **:E E**.

Regarding phrase-final tones, RaP and AM$^+$ assume that right edges of constituents license a variable number of unstarred tones, depending on postlexical, language-specific rules. Thus, for example, final rises are treated as two unstarred tones (**H H** %) when a slope change is observed; otherwise, in the case of monotonic rise, only one tone (**H** %) is warranted. As a further illustration of how RaP characterizes meaningful pitch range differences, take the *calling contour*, which entails a stepping down by a small pitch interval from one level-pitched stretch to another, as in *An-na-belle!* RaP repurposes ToBI's ! symbol to indicate a small pitch interval ([+small]). RaP's transcription for the calling contour is therefore **:E* E+ !L* +E %**.

- *Slope changes*: RaP codifies vertices corresponding to a slope change as tones. This allows a principled means of accounting for phenomena like ToBI's H- H% in cases when a shallow rise transitions on the last syllable to a steep rise. RaP allows the notation > for the upper edge of the pitch range or < for the lower edge of the pitch range. Pierrehumbert (1980) assumed uniformly that all intonation phrases end with two tones, a phrase accent and a boundary tone, that reduced phonetic transparency. RaP assumes that the number of edge tones may vary. This convention increases phonetic transparency because it is not necessary to assume that tones are present when there is no change in F0 slope. Note that this convention provides a principled account for the well-known problem of Greek prenuclear accents, which were problematic for ToBI*; this is because these Greek prenuclear accents show evidence of a slope change as their phonetic correlate, with a characteristic F0 valley and peak on the syllables preceding and following the accented syllable, respectively. RaP characterizes Greek prenuclear accents as **L+ !H* +H**, capturing the observed slope change (Arvaniti, Ladd, and Mennen 1998, 2000).

- *Sparse tonal representation*: Consistent with a sparse tonal representation, adjacent syntagmatic features are required to have different featural specifications. As a result, for example, when two rising intervals—[–same, +higher]—are adjacent to one another, one of them must be [-small] and the other [+small]. Phonologically, adjacent syntagmatic features of [+same] are thus banned for T_1 T_2 T_3. Phonetically, this corresponds to a slope change, with a tone—starred or unstarred—indicated at the locus of the slope change. As a consequence of these assumptions, there are no sequences like **E* +E**, **E* E+** or **E* E**, meaning that Pierrehumbert's (1980) assertion that strings of L* accents may give rise to a low, flat pitch is not supported in the present theory.

Conclusion

AM⁺ offers a simplified theory that accounts for a range of cross-linguistic tonal phenomena. Its approach is implemented with the RaP annotation system, which offers a phonetically transparent alternative to ToBI*. This phonetic transparency makes RaP a useful starting point for developing a phonologically principled International Prosodic Alphabet (IPrA) (Hualde and Prieto 2016).[22] RaP has been implemented as a full annotation system, with a publicly available set of interactive training materials[23] and a corpus of RaP-labeled speech (Breen et al. 2018). A large-scale study comparing annotation agreement between labelers trained in both the RaP and ToBI systems demonstrated RaP agreement levels that were equal to, and in some cases exceeded, agreement levels for ToBI (Breen et al. 2012). Finally, recent studies have successfully used the RaP system to accurately assess prosodic structure (Sharpe, Fogerty, and van Ouden 2017).

AM⁺ and RaP have already yielded new insights into metrical interactions between segmental and suprasegmental structures. Pierrehumbert (2000) noted that ToBI* failed to capture observed restrictions on the sequencing of tones—for example, the observation that tones tend to be repeated. Using the approach outlined in AM⁺ theory, Dilley and colleagues have experimentally demonstrated powerful perceptual constraints on metrical structure that can perceptually "garden-path" listeners into hearing different organizations of words (Breen et al. 2014; Dilley, Mattys, and Vinke 2010; Dilley and McAuley 2008; Morrill, Dilley, and McAuley 2014; Morrill et al. 2014).

In sum, in this chapter, Jun presents a useful outline for prosody researchers who are unfamiliar with ToBI*, but it only skimmed the surface with respect to problems behind ToBI*. Here, we have summarized several problematic aspects of the traditional AM theory that are irreconcilably part of ToBI*'s notations. We feel this is a critical juncture in time that will determine the usefulness of transcription choices to fields that stand to gain the most from consistent prosodic annotation. AM⁺ is a theory that retains the insights of traditional AM approaches that have stood the test of time, while affording new insights and considerable improvement in phonetic transparency. RaP and AM^+ are informed by more than forty years of research in phonetics, phonology, music cognition, and cognitive neuroscience. We hope that researchers will embrace paradigm change by moving toward AM^+ and a phonetically transparent system like RaP, in the interests of fostering further discovery in prosody research.

Notes

1. However, see Yi Xu and colleagues (Xu and Wang 2001; Xu and Xu 2005) for an opposing point of view.

2. Goldsmith's (1976) specific proposal was that H and L tones were based on features of [+/– high] and [+/– low]; H was [+high, -low], L was [-high, +low], and M was [+high, +low].

3. Following an a priori premise through to absurd illogical conclusions is an important part of philosophical scholarly enterprise. Philosophers routinely engage in thought experiments that involve alternative conceptualizations of reality that can lead to new insights (e.g., the Chinese room argument; Searle 1980). There is value in the philosophical traditions in the humanities by their permitting deeper understanding of what *is* by entertaining what *is not*. However, if applications of a priori reasoning result in unrealistic ideas about speech perception and production, the result will be to disconnect scholarly linguistic enterprises from science. The fact that unscientific approaches are common in linguistics (de Lacy 2014; Gibson and Fedorenko 2010) reflects well-known tensions between science- and humanities-oriented scholars who bump elbows in many linguistics departments.

4. The H + L* accent was later quietly "rescinded" from the English inventory by grouping it together with H* in MAE-ToBI and explicitly marking the lowering of a H* that follows another H* as a downstepped !H*. The idea of floating low tones being responsible for lowering of H—rather than some direct syntagmatic relationship—has never been retracted.

5. Mathematically, if $f: X \rightarrow Y$ is a set function from a collection of sets X to an ordered set Y, then f is said to be monotone if whenever $A \subseteq B$ as elements of X, $f(A) \leq f(B)$.

6. The obligatory contour principle (OCP), originally proposed by Leben (1973) to account for syntagmatic level pitch observed for sequences of paradigmatically specific lexical high tones, is described in AM^+ as the default assignment of the syntagmatic feature [+same] to sequences of lexically specified tones sharing a common paradigmatic tonal referent, *r*. Given that the OCP descriptively captures widespread phonological perceptual phenomena across languages (Berent, Shimron, and Vaknin 2001; Coetzee 2005; McCarthy 1986), we speculate that the OCP and the feature [+/– same] both reflect cognitive processes of attention and memory consolidation for processing sensory information that is the *same* versus *different* (Jones 1976; Large and Jones 1999).

7. Pitch targets associated with F0 elbows at right edges of flat stretches of pitch marking transition points to a rise or a fall are simply annotated **E** in RaP.

8. For example, in Mandarin, nearly every syllable is lexically specified for tone (Xu and Wang 2001), except for neutral tone syllables (Chen and Xu 2006; Lai and Dilley 2016). By contrast, in Chichewa, a Bantu language spoken mainly in Malawi, paradigmatic tonal specification appears to be much sparser (Myers 1998).

9. Dilley and Brown (2007) pointed out that the precise fix to the phonetic module proposed in Pierrehumbert and Beckman (1988) is to specify a rule that H tones cannot fall below adjacent L tones. Dilley and Brown argue that this rule-based systematicity is better captured by revisiting the idea of syntagmatic specifications as part of phonological representations.

10. We propose that pitch targets are associated with timing slots that specify locations of change in the velocity of pitch change over time. Viewed in this way, pitch targets encode the points of pitch *acceleration* as temporally coordinated with the metrical structure of speech utterances. The acoustic consequences of these pitch targets are predicted to roughly correspond to the second derivative of a function $y = f(x)$—that is $f''(x)$—where x is time and y is the F0 value. Then cases when $f'(x) = 0$ for a smoothed or stylized F0 contour correspond to F0 maxima and minima, whereas cases where $f''(x) = 0$ correspond to other slope changes, including elbows and changes in the steepness of a rise or fall, as approximated via the juncture point of two piecewise-linear F0 functions. An example of such a slope change is the +!H* tonal target in ToBI*'s H + !H* or a H- tonal target characterized as a slope change in Pierrehumbert's (1980) and ToBI*'s H-H%.

11. See Breen et al. (2012) for a conversion from MAE_ToBI to RaP. Converting from RaP to MAE_ToBI entails information loss.

12. RaP's **E** option appears to provide a solution to accounting for an interesting accentual distinction reported recently by Niebuhr and Hoekstra (2015) for North Frisian involving pointed versus plateau-shaped pitch accents.

13. The AM^{+} theory can be viewed as validating or mirroring the phonetically transparent INTSINT (International Transcription System for Intonation) approach by Daniel Hirst (Hirst and Di Cristo 1998), providing a theoretical framework that supports its insights.

14. Research within the AM framework in intonational phonology has sometimes oversimplified the debates regarding the nature of intonational representations and prematurely rejected syntagmatic representations (see, e.g., Arvaniti, Ladd, and Mennen 1998, 23). Although rise/fall approaches (e.g., 't Hart, Collie, and Cohen 1990) did not correct predict timing aspects of F0 curves (Arvaniti, Ladd, and Mennen 1998; Ladd 2008), these approaches did not suffer from the serious problems of underdetermined F0 contour shape from phonological primitives, which exists with the theories of Pierrehumbert (1980) and Pierrehumbert and Beckman (1988), as pointed out in Dilley and Brown (2007).

15. In Dilley (2005), the tonal referent was symbolized μ, the mathematical symbol for mean. We rename it r here to avoid confusion with moras.

16. In musical scale systems, each scale tone is defined by a relationship to a common paradigmatic referent pitch—termed the *tonic* in Western tonal music. This allows the notes to be "scaled" up or down—or instruments to be "tuned" up or down—creating completely different sets of absolute frequencies in hertz, while still allowing the melody (i.e., sequence of syntagmatically related pitches) to be recognized.

17. Indeed, musical scale systems have this property because they involve a common referent note. In Western music, the first scale note is called the tonic, and this note also names the key. For example, in the key of C, C is the tonic. When keys are changed in scale systems, the presence of a common referent note allows listeners to perceive the melody as constant even though the frequencies are shifted up or down (Dowling and Harwood 1986; Jones, Fay, and Popper 2010). In cases when there is no common referent note, as in atonal music, the pattern of ups and downs is the basis of the cognitive representation of the pitch sequence (Dowling and Fujitani 1971; Dowling and Harwood 1986), consistent with the primacy of syntagmatic features in the present theory. Note that musical intervals are perceived categorically (Siegel and Siegel 1977), as are lexical tones (Hallé, Chang, and Best 2004); thus, frequency ratios do not need to be exact to instantiate a tonal category. (See also Pfordresher et al. 2015.)

18. For example, Myers (1998) states that in Chichewa, a string of morphemes with low tone "is always realized unchanged with all low tones." He describes low tone as "phonologically inert, because it is simply the absence of tone" (367). This leads him to propose that low tone is underspecified in the surface representation. This description is consistent with a lexical paradigmatic specification that low tone in Chichewa is at the same level as *r*, T_L / *r* = [+same], which is the speaker's habitual (or mean) pitch phonetically.

19. There is considerable evidence that world musical systems are based on frequency ratios (Burns 1999; Perlman and Krumhansl 1996; Wright 2009). Dilley (2005) outlines an elaboration of the ideas presented here to account for how specific frequency ratios, or ranges of ratios, could become lexicalized. Fundamentally, we assume that F0 is a low-bandwidth channel in an information-theoretic sense (Shannon 1948), limiting the number of paradigmatic tonal contrasts that can be transmitted through it. In the speech-to-song transformation, a phrase that is repeated several times shifts perceptually to being heard as sung (Falk, Rathcke, and Dalla Bella 2014; Tierney, Patel, and Breen 2018). This phenomenon supports the contention that the pitches in speech are subject to similar cognitive organizing principles as in music.

20. Syntagmatic featural specifications at higher AM^+ grid tiers constrain the possible steps among tones on lower grid tiers. Tones that propagate to higher grid tiers are the more "important" ones; though they are nonadjacent in time, they are heard to form a cohesive syntagmatic structure. For example, in J. S. Bach's Toccata and Fugue in D Minor, the solo unaccompanied allegro passage involves an alternation between low and high notes; the low notes in this passage are metrically prominent and are heard to form a coherent melody, even though they are nonadjacent in the note sequence. Dilley (2005) proposes the multiplicative property to capture how the syntagmatic relationships among tones on different AM^+ grid tiers are constrained relative to one another. Generalizing Dilley's formulation, this property says that for two tones, T_n and T_N for $n, n+1, \ldots, N$ which are adjacent on a higher grid tier, M + 1, the syntagmatic features of tones subtended by T_n and T_N constrains the sequence of steps at the next lower grid tier, M, such that T_N / T_n must equal $(T_{n+1} / T_n) \cdot (T_{n+2} / T_{n+1}) \cdot \ldots (T_N / T_{N-1})$. The notation conveys abstract relationships, but it also has a direct mathematical interpretation, because frequency ratios in music are multiplicative (Wright 2009). Note that frequency ratios do not need to be exact to be heard as instances of a musical category (Siegel and Siegel 1977).

21. Other phonetic differences have sometimes been found between accents realizing focus differences, such as a later peak for L + H*. RaP would capture an audibly late peak that occurred within the accented syllable as an extra unstarred tone, leading to a variant contour: **L+ H* !H(+)** . . .

22. See also Hirst (chapter 3, this volume) and Hirst and Di Cristo (1998).

23. Available at http://tedlab.mit.edu/tedlab_website/RaPHome.html.

References

Arvaniti, A., D. R. Ladd, and I. Mennen. 1998. "Stability of Tonal Alignment: The Case of Greek Prenuclear Accents." *Journal of Phonetics* 26:3–25.

Arvaniti, A., D. R. Ladd, and I. Mennen. 2000. "What Is a Starred Tone? Evidence from Greek." In *Papers in Laboratory Phonology V*, edited by M. B. Broe and J. B. Pierrehumbert, 119–130. Cambridge: Cambridge University Press.

Barnes, J., N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel. 2012. "Tonal Center of Gravity: A Global Approach to Tonal Implementation in a Level-Based Intonational Phonology." *Laboratory Phonology* 3:337–383.

Beckman, M., and J. Hirschberg. 1994. "The ToBI Annotation Conventions." Ohio State University, Columbus, Ohio. https://web.archive.org/web/20130218081131/http://www.ling.ohio-state.edu/~tobi/.

Beckman, M., J. Hirschberg, and S. Shattuck-Hufnagel. 2005. “The Original ToBI System and the Evolution of the ToBI Framework.” In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun, 9–54. New York: Oxford University Press.

Berent, I., J. Shimron, and V. Vaknin. 2001. “Phonological Constraints on Reading: Evidence from the Obligatory Contour Principle.” *Journal of Memory and Language* 44:644–665.

Bishop, J., and P. A. Keating. 2012. “Perception of Pitch Location within a Speaker’s Range: Fundamental Frequency, Voice Quality, and Speaker Sex.” *Journal of the Acoustical Society of America* 132:1100–1112.

Breen, M., L. C. Dilley, M. Brown, and E. Gibson. 2018. Rhythm and Pitch (RaP) Corpus. LDC2018S04. Philadelphia: Linguistic Data Consortium.

Breen, M., L. C. Dilley, J. Kraemer, and E. Gibson. 2012. “Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch).” *Corpus Linguistics and Linguistic Theory* 8:277–312. https://doi.org/10.1515/cllt-2012-0011.

Breen, M., L. C. Dilley, J. D. McAuley, and L. D. Sanders. 2014. “Auditory Evoked Potentials Reveal Early Perceptual Effects of Distal Prosody on Speech Segmentation.” *Language, Cognition and Neuroscience* 29:1132–1146. https://doi.org/10.1080/23273798.2014.894642.

Breen, M., E. Fedorenko, M. Wagner, and E. Gibson. 2010. “Acoustic Correlates of Information Structure.” *Language and Cognitive Processes* 25:1044–1098.

Bregman, A. S. 1994. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Burnett, T. A., M. B. Freedland, C. R. Larson, and T. C. Hain. 1998. “Voice F0 Responses to Manipulations in Pitch Feedback.” *Journal of the Acoustical Society of America* 103:3153–3161.

Burns, E. M. 1999. “Intervals, Scales, and Tuning.” In *The Psychology of Music*, 2nd ed., edited by D. Deutsch, 215–264. San Diego: Academic Press.

Chen, S., H. Liu, Y. Xu, and C. R. Larson. 2007. “Voice F0 Responses to Pitch-Shifted Voice Feedback during English Speech.” *Journal of the Acoustical Society of America* 121:1157–1163.

Chen, Y., and Y. Xu. 2006. “Production of Weak Elements in Speech: Evidence from F0 Patterns of Neutral Tone in Standard Chinese.” *Phonetica* 63:47–75.

Clements, G. N., and J. Goldsmith, eds. 1984. *Autosegmental Studies in Bantu Tone*. Dordrecht, the Netherlands: Foris.

Coetzee, A. 2005. “The Obligatory Contour Principle in the Perception of English.” In *Prosodies: With Special Reference to Iberian Languages*, edited by S. Frota, M. Vigário, and M. J. Freitas, 223–246. Berlin: Walter de Gruyter.

Cole, J. 2015. “Prosody in Context: A Review.” *Language, Cognition and Neuroscience* 30:1–31.

Cutler, A., D. Dahan, and W. van Donselaar. 1997. “Prosody in the Comprehension of Spoken Language: A Literature Review.” *Language and Speech* 40:141–201.

d’Alessandro, C., and P. Mertens. 1995. “Automatic Pitch Contour Stylization Using a Model of Tonal Perception.” *Computer Speech and Language* 9:257–288.

D’Imperio, M. 2000. “The Role of Perception in Defining Tonal Targets and Their Alignment.” PhD diss., Ohio State University.

D’Imperio, M., B. Gili Fivela, and O. Niebuhr. 2010. “Alignment Perception of High Intonational Plateaux in Italian and German.” Paper presented at the International Conference on Speech Prosody, Chicago, IL, USA, May.

de Lacy, P. 2002. “The Interaction of Tone and Stress in Optimality Theory.” *Phonology* 19:1–32.

de Lacy, P. 2014. Evaluating Evidence for Stress Systems.” In *Word Stress: Theoretical and Typological Issues*, edited by H. v. d. Hulst, 149–193. Cambridge: Cambridge University Press.

del Giudice, A., R. Shosted, K. Davidson, M. Salihie, and A. Arvaniti. 2007. “Comparing Methods for Locating Pitch `Elbows.’” In *Proceedings of the Sixteenth International Congress of Phonetic Sciences*, http://www.icphs2007.de/conference/Papers/1283/index.html.

Dilley, L. C. 2005. “The Phonetics and Phonology of Tonal Systems.” PhD diss., MIT.

Dilley, L. C. 2010. “Pitch Range Variation in English Tonal Contrasts Is Continuous, Not Categorical.” *Phonetica* 67:63–81.

Dilley, L. C., and M. Brown. 2005. *The RaP (Rhythm and Pitch) Labeling System, Version 1.0*. East Lansing: Michigan State University. http://speechlab.cas.msu.edu/rap-system.htm.

Dilley, L. C., and M. Brown. 2007. “Effects of Pitch Range Variation on F0 Extrema in an Imitation Task.” *Journal of Phonetics* 35:523–551.

Dilley, L. C., and C. Heffner. 2013. “The Role of f0 Alignment in Distinguishing Intonation Categories: Evidence from American English.” *Journal of Speech Sciences* 3:3–67.

Dilley, L. C., and J. D. McAuley. 2008. Distal Prosodic Context Affects Word Segmentation and Lexical Processing.” *Journal of Memory and Language* 59:294–311. https://doi.org/10.1016/j.jml.2008.06.006.

Dilley, L. C., D. R. Ladd, and A. Schepman. 2005. “Alignment of L and H in Bitonal Pitch Accents: Testing Two Hypotheses.” *Journal of Phonetics* 33:115–119.

Dilley, L. C., S. Mattys, and L. Vinke. 2010. “Potent Prosody: Comparing the Effects of Distal Prosody, Proximal Prosody, and Semantic Context on Word Segmentation.” *Journal of Memory and Language* 63:274–294. https://doi.org/10.1016/j.jml.2010.06.003.

Dowling, W. J., and D. S. Fujitani. 1971. “Contour, Interval, and Pitch Recognition in Memory for Melodies.” *Journal of the Acoustical Society of America* 49:524–531.

Dowling, W. J., and D. L. Harwood. 1986. *Music Cognition*. Orlando: Academic Press.

Falk, S., T. Rathcke, and S. Dalla Bella. 2014. “When Speech Sounds like Music.” *Journal of Experimental Psychology: Human Perception and Performance* 40:1491–1506.

Féry, C., and M. Krifka. 2008. “Information Structure: Notional Distinctions, Ways of Expression.” In *Unity and Diversity of Languages*, edited by P. v. Sterkenburg, 123–136. Amsterdam: John Benjamin.

Fitzroy, A. B., and M. Breen. 2019 “Metric Structure and Rhyme Predictability Modulate Speech Intensity During Child-Directed and Read-Alone Productions of Children’s Literature.” *Language and Speech* 63:292–305. https://doi.org/10.1177/0023830919843158.

Gibson, E., and E. Fedorenko. 2010. “Weak Quantitative Standards in Linguistics Research.” *Trends in Cognitive Science* 14:233–234.

Goldsmith, J. 1976. “Autosegmental Phonology.” PhD diss., MIT.

Greenberg, S., H. Carvey, and L. Hitchcock. 2002. “The Relationship between Stress Accent and Pronunciation Variation in Spontaneous American English Discourse.” In *Proceedings of the ISCA Workshop on Prosody and Speech Processing*. https://www.isca-speech.org/archive_open/prosody_2001/prsr_009.html.

Grice, M. 1995. “Leading Tones and Downstep in English.” *Phonology* 12:183–233.

Grice, M., D. R. Ladd, and A. Arvaniti. 2000. "On the Place of Phrase Accents in Intonational Phonology." *Phonology* 17:143–186.

Guenther, F. H. 2016. *Neural Control of Speech*. Cambridge, MA: MIT Press.

Guenther, F. H., and G. Hickok. 2015. "Role of the Auditory System in Speech Production." In *Handbook of Clinical Neurology*, edited by M. J. Aminoff, F. Boller, and D. Swaab, 161–175. Amsterdam: Elsevier.

Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Halle, M., and W. Idsardi. 1995. "General Properties of Stress and Metrical Structure." In *The Handbook of Phonological Theory*, edited by J. A. Goldsmith, 403–441. Oxford: Blackwell.

Hallé, P. A., Y.-C. Chang, and C. T. Best. 2004. "Identification and Discrimination of Mandarin Chinese Tones by Mandarin Chinese vs. French Listeners." *Journal of Phonetics* 32 (3): 395–421.

Hannon, E. E., and L. J. Trainor. 2007. "Music Acquisition: Effects of Enculturation and Formal Training on Development." *Trends in Cognitive Science* 11:466–472.

Hayes, B. 1995. *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.

Hirst, D., and A. Di Cristo. 1998. *Intonation Systems: A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

Honorof, D. N., and D. H. Whalen. 2005. "Perception of Pitch Location within a Speaker's F0 range." *Journal of the Acoustical Society of America* 117:2193–2200.

House, D. 1990. *Tonal Perception in Speech*. Lund, Sweden: Lund University Press.

Hualde, J. H., and P. Prieto. 2016. "Towards an International Prosodic Alphabet (IPrA)." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7:1–25.

Hutchins, S., and I. Peretz. 2011. "Perception and Action in Singing." In *Progress in Brain Research*, edited by A. M. Green, C. E. Chapman, J. F. Kalaska, and F. Lepore, 103–118. Amsterdam: Elsevier.

Hyman, L. H., and R. G. Schuh. 1974. "Universals of Tone Rules: Evidence from West Africa." *Linguistic Inquiry* 5:81–115.

Hyman, L. M. 1993. "Register Tones and Tonal Geometry." In *The Phonology of Tone: The Representation of Tonal Register*, edited by H. van der Hulst and K. Snider, 75–108. Berlin: Mouton de Gruyter.

Inkelas, S., and W. R. Leben. 1990. "Where Phonology and Phonetics Intersect: The Case of Hausa Intonation." In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. Beckman, 17–34. Cambridge: Cambridge University Press.

Inkelas, S., W. R. Leben, and M. Cobler. 1986. "The Phonology of Intonation in Hausa." Paper presented at the Sixteenth Annual Meeting of the North East Linguistic Society, Amherst, MA, October.

Jakobson, R., G. Fant, and M. Halle. 1952. *Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press.

Jones, M. R. 1976. "Time, Our Lost Dimension: Toward a New Theory of Perception, Attention, and Memory." *Psychological Review* 83:323–355. https://doi.org/10.1037/0033-295X.83.5.323.

Jones, M. R., R. Fay, and A. Popper, eds. 2010. *Music Perception*. New York: Springer.

Katz, J., and E. Selkirk. 2011. "Contrastive Focus vs. Discourse—New: Evidence from Phonetic Prominence in English." *Language: Journal of the Linguistic Society of America* 87:771–816.

Knight, R.-A. 2008. "The Shape of Nuclear Falls and Their Effect on the Perception of Pitch and Prominence: Peaks vs. Plateaux." *Language and Speech* 51:223–244.

Knight, R.-A., and F. Nolan. 2006. "The Effect of Pitch Span on Intonational Plateaux." *Journal of the International Phonetic Association* 36:21–38.

Ladd, D. R. 1986. "Intonational Phrasing: The Case for Recursive Prosodic Structure." *Phonology Yearbook* 3:311–340.

Ladd, D. R. 1988. "Declination 'Reset' and the Hierarchical Organization of Utterances." *Journal of the Acoustical Society of America* 84:530–544.

Ladd, D. R. 2008. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Ladd, D. R., and A. Schepman. 2003. "'Sagging Transitions' between High Accent Peaks in English: Experimental Evidence." *Journal of Phonetics* 31:81–112.

Lai, W., and L. C. Dilley. 2016. "Cross-Linguistic Generalization of the Distal Rate Effect: Speech Rate in Context Affects Whether Listeners Hear a Function Word in Chinese Mandarin." *Proceedings of the International Conference on Speech Prosody* 8:1124–1128.

Large, E. W., and M. R. Jones. 1999. "The Dynamics of Attending: How People Track Time-Varying Events." *Psychological Review* 106:19–159. https://doi.org/10.1037/0033-295X.106.1.119.

Leben, W. R. 1973. "Suprasegmental Phonology." PhD diss., MIT.

Lee, C.-Y. 2009. "Identifying Isolated, Multispeaker Mandarin Tones from Brief Acoustic Input: A Perceptual and Acoustic Study." *Journal of the Acoustical Society of America* 125:1125–1137.

Liberman, M. 1975. *The Intonation System of English*. PhD diss., MIT.

Liberman, M., and J. Pierrehumbert. 1984. "Intonational Invariance under Changes in Pitch Range and Length." In *Language Sound Structure*, edited by M. Aronoff and R. Oerhle, 157–233. Cambridge, MA: MIT Press.

Manfredi, V. 1993. "Spreading and Downstep: Prosodic Government in Tone Languages." In *The Phonology of Tone: The Representation of Tonal Register*, edited by H. van der Hulst and K. Snider, 133–184. Berlin: Mouton de Gruyter.

McCarthy, J. 1986. "OCP Effects: Gemination and Antigemination." *Linguistic Inquiry* 17:207–263.

Mertens, P. 2004. "The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model." Paper presented at the International Conference on Speech Prosody, Nara, Japan, March.

Monelle, R. 2014. *Linguistics and Semiotics in Music*. London: Routledge.

Morrill, T., L. C. Dilley, and J. D. McAuley. 2014. "Prosodic Patterning in Distal Speech Context: Effects of List Intonation and f0 Downtrend on Perception of Proximal Prosodic Structure." *Journal of Phonetics* 46:68–85.

Morrill, T., L. C. Dilley, J. D. McAuley, and M. A. Pitt. 2014. "Distal Rhythm Influences Whether or Not Listeners Hear a Word in Continuous Speech: Support for a Perceptual Grouping Hypothesis." *Cognition* 131:69–74. https://doi.org/10.1016/j.cognition.2013.12.006.

Myers, S. 1998. "Surface Underspecification of Tone in Chichewa." *Phonology* 15:367–392.

Niebuhr, O. 2007a. *Perzeption un kognitive Verarbeitung der Sprechmelodie: Theoretische Grundlagen und empirische Untersuchungen*. Berlin: Walter de Gruyter.

Niebuhr, O. 2007b. "The Signalling of German Rising-Falling Intonation Categories: The Interplay of Synchronization, Shape, and Height." *Phonetica* 64:174–193.

Niebuhr, O., and J. Hoekstra, J. 2015. "Pointed and Plateau-Shaped Pitch Accents in North Frisian." *Laboratory Phonology* 6:433–468.

Ning, L. H., C. Shih, and T. M. Loucks. 2014. "Mandarin Tone Learning in L2 Adults: A Test of Perceptual and Sensorimotor Contributions." *Speech Communication* 63:55–69.

O'Connor, R. J., and G. F. Arnold. 1973. *Intonation of Colloquial English*. Bristol, UK: Longman.

Odden, D. 1995. "Tone: African Languages." In *The Handbook of Phonological Theory*, edited by J. Goldsmith, 444–475. Oxford: Blackwell.

Patel, A. D. 2010. *Music, Language, and the Brain*. New York: Oxford University Press.

Patel, R., C. Niziolek, K. Reilly, and F. H Guenther. 2011. "Prosodic Adaptations to Pitch Perturbation in Running Speech." *Journal of Speech, Language, and Hearing Research* 54:1051–1059.

Peng, G., C. Zhang, H.-Y. Zheng, J. Minnett, and, W. S.-Y. Wang. 2012. "The Effect of Intertalker Variations on Acoustic-Perceptual Mapping in Cantonese and Mandarin Tone Systems." *Journal of Speech, Language, and Hearing Research* 55:579–595.

Perlman, M., and C. L. Krumhansl. 1996. "An Experimental Study of Internal Standards in Javanese and Western Musicians." *Music Perception* 14:95–116.

Pfordresher, P. Q., S. M. Demorest, S. Dalla Bella, S. Hutchins, P. Loui, J. Rutkowski, and G. F. Welch. 2015. "Theoretical Perspectives on Singing Accuracy: An Introduction to the Special Issue on Singing Accuracy (Part 1)." *Music Perception: An Interdisciplinary Journal* 32:227–231.

Pierrehumbert, J. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pierrehumbert, J. 2000. "Tonal Elements and Their Alignment." In *Prosody: Theory and Experiment*, edited by M. Horne, 11–36. Dordrecht, the Netherlands: Kluwer.

Pierrehumbert, J., and M. Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pierrehumbert, J., and S. A. Steele. 1989. "Categories of Tonal Alignment in English." *Phonetica* 46:181–196.

Prieto, P., M. D'Imperio, and B. Gili Fivela. 2005. "Pitch Accent Alignment in Romance: Primary and Secondary Associations with Metrical Structure." *Language: Journal of the Linguistic Society of America* 48:359–396.

Rice, K. D. 1987. "Metrical Structure in a Tone Language: The Foot in Slave (Athapaskan)." Paper presented at the Twenty-Third Annual Regional Meeting of the Chicago Linguistics Society.

Searle, J. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3:417–457.

Shannon, C. E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27:379–423, 623–656.

Sharpe, V., D. Fogerty, and D.-B. van Ouden. 2017. “The Role of Fundamental Frequency and Temporal Envelope in Processing Sentences with Temporary Syntactic Ambiguities.” *Language and Speech* 60:399–426.

Siegel, J. A., and W. Siegel. 1977. “Categorical Perception of Tonal Intervals: Musicians Can’t Tell *Sharp* from *Flat*.” *Perception and Psychophysics* 21:399–407.

Snider, K. 1999. *The Geometry and Features of Tone*. Dallas: Summer Institute of Linguistics and the University of Texas at Arlington Publications in Linguistics.

’t Hart, J., R. Collier, and A. Cohen. 1990. *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press.

Tierney, A., A. D. Patel, and M. Breen. 2018. “Repetition Enhances the Musicality of Speech and Tone Stimuli to Similar Degrees.” *Music Perception: An Interdisciplinary Journal*. 35 (5): 573–578.

Truckenbrodt, H. 2002. “Upstep and Embedded Register Levels.” *Phonology* 19:77–120.

Vos, P., and J. Troost. 1989. “Ascending and Descending Melodic Intervals: Statistical Findings and Their Perceptual Relevance.” *Music Perception* 6:383–396.

Welby, P. 2006. “French Intonational Structure: Evidence from Tonal Alignment.” *Journal of Phonetics* 34:343–371.

Williams, E. S. (1971) 1976. “Underlying Tone in Margi and Igbo.” *Linguistic Inquiry* 7:463–484. Manuscript was written in 1971.

Wong, P. C. M., and R. L. Diehl. 2003. “Perceptual Normalization for Inter- and Intra-Talker Variation in Cantonese Level Tones.” *Journal of Speech, Language, and Hearing Research* 46:413–421.

Wright, D. 2009. *Mathematics and Music*. Providence, RI: American Mathematical Society.

Xu, Y., and C. Xu. 2005. “Phonetic Realization of Focus in English Declarative Intonation.” *Journal of Phonetics* 33:159–197.

Xu, Y., and W. E. Wang. 2001. “Pitch Targets and Their Realization: Evidence from Mandarin Chinese.” *Speech Communication* 33:319–337.

Zec, D. 1999. “Footed Tones and Tonal Feet: Rhythmic Constituency in a Pitch-Accent Language.” *Phonology* 16:225–271.

Author Response to the Commentary: ToBI Is Not Designed to Be Phonetically Transparent

Sun-Ah Jun

Introduction

My essay for this chapter introduced how the tones and break indices (ToBI) transcription system works, and by referring to the ToBI systems of various languages, it discussed the strengths and challenges associated with the ToBI system and some of the recent developments made in response to such challenges. I was asked by the volume editors to present the ToBI system to relative newcomers to the field of prosody so that they can decide which transcription method is the one they are looking for, given their individual goals. I was also asked not to introduce the autosegment-metrical (AM) theory in depth, because this is addressed by Arvaniti (chapter 1, this volume). Therefore, my chapter was not about AM theory in general, but about ToBI in particular, and so AM theory was only described to the extent necessary to understand the ToBI system.

In their commentary, Dilley and Breen argue that AM theory has several serious problems, leading to limitations on ToBI*'s (= all ToBI-like annotation systems) value as a scientific tool. They summarized AM theory before criticizing it and introduced what they call an enhanced AM theory (AM$^+$). Their chief criticism of AM theory was that it does not include syntagmatic tonal relationships in phonology, which they also cite as a source of problems with ToBI*'s phonetic transparency and consistency (see "Enduring Insights from over Forty Years of Traditional AM Theory").

To fix this perceived problem, they included in AM$^+$ theory both syntagmatic and paradigmatic relationships of tones in the phonology and presented the rhythm and pitch (RaP) prosodic transcription system, which is an instantiation of their AM$^+$ theory. In this response, I will argue that Dilley and Breen's criticisms of (standard) AM theory and ToBI* are not fully justified, and in so doing, I will also highlight how AM theory and ToBI* are quite different from AM$^+$ theory and the RaP system.

The AM Theory versus the AM Model of English Intonation

In their commentary, Dilley and Breen list three main problems with AM theory that they claim are due to the assumption that the phonology lacks syntagmatic restrictions on tones: (i) complex phonetic rules and mechanisms for tone scaling that do not work, (ii) inconsistencies in mapping pitch accents to F0 events, and (iii) complications when F0 curves correspond to phonetic interpolation versus tones. They claimed that these problems in AM theory then translate to problems for systems like ToBI*, particularly the use of such systems for research on prosodic typology (see "Strictly Paradigmatic Phonological Representations"). However, these problems are mainly based on the AM model of English intonation proposed in Pierrehumbert (1980), and it should be noted that the phonological categories and phonetic rules proposed for English intonation in Pierrehumbert (1980) and later works by Pierrehumbert and colleagues to explain surface F0 variations are specific to English, and thus do not apply equally to AM models of the intonational phonology of other, typologically different languages. What has been adopted from the AM model of English intonation by researchers working on models of other languages are instead the basic principles and assumptions of AM theory, such as:

- Intonation contours are analyzed as a linear sequence of two tonal targets, H (high) and L (low), and their combinations (e.g., bitones such as H + L or L + H).
- Unlike a single tone, a bitone has a close relationship between two adjacent tonal targets aligned to the hosting syllable or mora.
- Tones can have fundamentally different functions (marking prominence relations among words via pitch accents or marking the hierarchical structure of an utterance via boundary tones) and can change pragmatic or semantic meanings or cue syntactic structures.
- F0 heights of tones as well as the alignment of the tone-text matter are used in deciding a tonal category. If any of these differences changes the meaning or the prosodic structure of the utterance, that is, if any of them is distinctive, it should be coded into a tone.
- Pitch range can change throughout an utterance, reflecting the degree of prominence, the informational status of a word, the pragmatic and discourse meaning of a phrase, and the prosodic phrasing of the utterance.
- Prosodic phrasing can be defined by pitch-range reset as well as by the degree of phrase-final lengthening and the presence of a boundary tone.

MAE_ToBI versus ToBI*

It is also important to highlight that English ToBI, or more specifically, MAE_ToBI (Mainstream American English tones and break indices), is not a direct instantiation of the AM model of English intonation proposed by Pierrehumbert (1980). As Beckman, Hirschberg, and Shattuck-Hufnagel (2005) noted, a great deal of literature preceding Pierrehumbert (1980) and Beckman and Pierrehumbert (1986) contributed to the development of the MAE_ToBI. Furthermore, when a ToBI system is developed for a language whose prosodic system is quite different from English (for example, languages like Korean that lack both lexical stress and intonational pitch accent), some of the main principles proposed in MAE_ToBI had to be modified or extended, as mentioned in my essay. Therefore, criticisms of language-specific implementations like MAE_ToBI do not straightforwardly apply to ToBI* generally, and Dilley and Breen's argumentation is rather misleading in this regard.

Difficulties with Using ToBI* for Studying Prosodic Typology

Dilley and Breen claim that AM theory's lack of syntagmatic phonological restrictions on tones has led to difficulties in using ToBI* for prosodic typology, citing Hualde and Prieto (2016). However, the difficulties that Hualde and Prieto discuss are not at all specific to ToBI*, nor are they related to a lack of syntagmatic phonological specifications. Rather, the difficulties that emerge are due to ToBI*'s status as a phonological transcription system. In fact, this was one of the primary challenges to the ToBI system that is highlighted in my essay: there are inherent limits to the use of such a system for the purposes of typological description. Phonological contrasts are by nature language-specific, and thus the same tonal label can "mean" different things across languages—just as is seen in typological study of segmental sound patterns. Typological study therefore requires consideration of both phonological and phonetic components of the target languages' sound systems, and difficulties will arise if either is ignored. To study prosodic typology, the AM approach to intonation would provide evidence primarily regarding phonological contrasts, and thus we would need to look elsewhere for evidence bearing on the relevant aspects of phonetic realization. Such evidence could come from acoustic data, but, as noted in my essay, it could also come from phonetically transparent labels of the sort being developed for an IPrA (international prosodic alphabet). As a phonological transcription system, ToBI* is neither intended to be nor expected to be phonetically transparent, and so its lack of phonetic transparency is a poor basis for criticism. Indeed, the fact that it is possible to use some ToBI labels in a phonetically transparent way, or to mix phonological and phonetic labels, was cited as one of the weaknesses of ToBI* in my essay. If a transcription system is phonetically transparent—that is, its labels pick out nondistinctive surface prosodic properties—it cannot and should not be used as a phonological transcription system.

Syntagmatic Tonal Relationships in AM Theory

There is no question that we need to consider both paradigmatic and syntagmatic properties of tones in the study of tone and intonation. As rightly pointed out in Dilley and Breen's commentary, Pierrehumbert (1980) and Pierrehumbert and Beckman (1988) did not ignore syntagmatic relationships for underlying tonal representations; although Pierrehumbert (1980) and later works on the AM model of English intonation (including Beckman and Pierrehumbert 1986) defined the L and H tonal targets paradigmatically by referring to the pitch range of a speaker at a given point in an utterance, they did include syntagmatic relationships among tones in phonology. This is, for example, evidenced in their inventory of bitonal pitch accents (e.g., L+H*, L*+H, H+L*) and the addition of an H*+L ("floating L" tone) to capture the downstep relation between two adjacent H tones. In fact, their AM model of English was not the first model to codify the importance of syntagmatic tonal relations to English intonation, as the matter was also central to earlier debate over levels versus configurations (see discussion in Bolinger 1951). Pierrehumbert (1980) and later works made their approach distinct from the earlier research by proposing that underlying tonal representations were simpler and more abstract (by allowing only two tonal levels, instead of four as in Trager and Smith 1951) and by deriving the various nondistinctive surface realizations of underlying tones in the phonetic component of the grammar. As mentioned in my essay, the surface F0 variations, including the relative F0 height of a sequence of singleton pitch accents, are explained by the degree of prominence, a tone's lexical status, and the type of tone sequence. Importantly, these factors cover both paradigmatic (e.g., the top reference line of nuclear H*'s pitch range can be higher than that of non-nuclear H*'s) and syntagmatic relationships (e.g., H% is realized higher than the preceding H-; L% is lower than the L of L+H*) of tones. Thus, a central tenet of AM theory is to allow only distinctive tonal events in the underlying representation. Any f0 events that are nondistinctive, predictable, or unmarked (e.g., a speaker's medium/unmarked F0 level at the onset of an utterance) are outside of the phonological component of the grammar.

Dilley and Breen criticized AM theory for not including syntagmatic tonal relationships in phonology, stating in their "Summary" section: "Fortunately, ToBI* systems have been used by communities of scholars as though syntagmatic tonal relationships are part of the phonology, even though they are not. For example, scholars have annotated L + H* as a low valley plus a rising pitch." However, as mentioned herein, the AM model of English intonation proposed in Pierrehumbert (1980) and revised in Beckman and Pierrehumbert (1986) explicitly treated syntagmatic tonal relationships as part of the phonology and specified how bitonal pitch accents differ from singleton pitch accents. For example, Beckman and Pierrehumbert specified that "the difference between L* + H and H* involves contrast not only in the timing of peak but also in the F0 level immediately preceding the peak. The L* + H accent has a valley on the stressed syllable which is as low as that for any L* accent whereas the H* accent has no such valley" (259). The MAE_ToBI manual (Beckman and Ayers-Elam 1997) further illustrates in section 2.2 how L + H* is acoustically different from H* or from a sequence of L* and H*. Section 2.6 of the manual also illustrates how L + H* is acoustically different from L* + H, and the alignment difference is distinctive in English, demonstrating that the timing of the F0 rise relative to the accented syllable can be also phonological. Therefore, it is natural and expected that ToBI systems of other languages have included syntagmatic tonal relationships as part of phonology.

AM$^+$ Theory and the RaP Prosodic Transcription System

The details of AM$^+$ theory and the RaP prosodic transcription system, as described in Dilley and Breen's commentary, suggest that the goal of AM$^+$ is to make a prosodic transcription phonetically transparent and to provide a theory for such transcriptions. Indeed, "phonetically transparent" is explicit both in the commentary's title and their note 12 about the AM$^+$ theory validating the INTSINT (International Transcription System for Intonation) approach (Hirst and Di Cristo 1998, and Hirst, chapter 3, this volume). INTSINT is known as a narrow phonetic transcription of F0 contours.

To achieve such a goal, AM$^+$ and RaP are designed to transcribe F0 changes (turning points and slope changes) in terms of symbols representing syntagmatic tonal relationships. However, it seems that phonetic transparency is enhanced in RaP not because the AM$^+$ prioritizes syntagmatic tonal relationships (as Dilley and Breen emphasize throughout their commentary), but because the AM$^+$ theory made its "phonological" component very rich and its categories very fine-grained. The phonological component in their theory includes three features capturing syntagmatic tonal relationships—[same], [higher], and [small][1]—as well as other categories (with a few restrictions on their combination). Because the feature values of [higher] or [small] are specified only with [–same], combinations of these three features will generate five types of syntagmatic relationships between two adjacent tones, represented by five symbols in RaP: **H**, **!H**, **E**, **!L**, **L**. (Following the convention used in the commentary, I will mark the RaP symbols in bold.) **H** is for a tone that has the feature specifications [–same, +higher], **L** for those with [–same, –higher], **E** for [+same], **!H** for [–same, +higher, +small], and **!L** for [–same, –higher, +small]. In addition to these five symbols for syntagmatic F0 targets, RaP includes the **+** symbol to represent the association of an unstarred tone (on a non-prominent syllable) to an adjacent starred tone (on a prominent syllable). **+** can be put on the right side or to the left side of any unstarred tone symbol (meaning an unstarred tone can come before or after a starred tone), theoretically creating multiple combinations (e.g., **H* +L**, **H* +E**, **H+ L***, **H+ E***, **L* +H**, **L+ H***, **L* +E**, **E+ L***, **E* +H**, **!H* +H**, **H+**

!L*). Furthermore, RaP posits three phrase-initial phonological tones, designated by :, namely, **:L H** (omitting + and *) for a rise, **:H L** for a fall, and **:E E** for a level pitch, and it does not restrict the number of phrase-final tones (e.g., **H%, HH%**). Having all of these categories, AM$^+$/RaP would easily allow for the transcription of all possible F0 contours in English or any other language at the phonetic level. At the phonological level, however, it would predict many more tonal contrasts than any one language could have, and at the same time, it would generate many impossible or unobserved F0 contours, suggesting that AM$^+$ has a problem of overgeneration as a phonological model.

To illustrate their system's capacity for phonetic transparency, Dilley and Breen proposed RaP tonal labels for a few cases that pose well-known challenges for pitch accent categories in ToBI. First, they proposed labeling the H* versus L + H* in MAE_ToBI as **L+ !H*** versus **L+ H***, respectively, in RaP. This would allow their system to capture the difference between a small rise to a peak for H* versus a large rise to a peak for L + H* in English. However, these two RaP labels misrepresent the tonal contrast of the two pitch-accent types in English. While it is true that F0 is generally higher for L + H* than H*, it is the presence or absence of the preceding L tone target, not the height of the H target, which is distinctive in English; L + H* requires a low F0 target immediately before the H* target, but H* does not. For H*, F0 rises gradually toward the H* target and no L target should be adjacent to it. In fact, F0 does not have to rise to a peak at all to be labeled as H*; an H* can be preceded by a high F0 (e.g., a "hat pattern"). Labeling ToBI's H* as **L+ !H*** in RaP further suggests that the symbol **!H** refers to a small, local F0 rise toward a single high target, independent of the F0 level of the preceding H tone. That is, RaP would allow **!H** to occur as the first H tone of a phrase, because **!H** does not indicate that the H target is lower than a preceding H as it does in MAE_ToBI. A **!L** symbol is used in RaP to represent an F0 target slightly lower than the preceding tone. Thus the **!H** symbol does not serve the purpose of capturing the global pitch range reduction applied to all post-!H tones as it does in MAE_ToBI. And it is not clear how this phenomenon of global pitch-range reduction (beyond the window of syntagmatic relation between two adjacent tonal targets) can be represented, even on higher grid tiers, in RaP. While RaP could mark lowered pitch range with a specific diacritic before each affected tone, this would not really capture the global effect.

The idea of using **!H** to denote a local tone level for a small rise to a peak compared to a large rise to a peak is also exemplified in their proposal for Greek prenuclear pitch accents, for which they assign **L+ !H* +H**. One might think that this label closely matches what is observed in the surface F0 contour, namely, the low F0 valley before the accented syllable whose F0 is slightly lower than that of the postaccented syllable. However, this would work only if the label represents paradigmatic tonal height, which is not the case in RaP. Furthermore, given the phonological constraints on assigning pitch accent status in the AM$^+$ theory—a metrically prominent syllable without any local pitch change is not considered a pitch accent—the **L+ !H* +H** label in RaP would not match the described surface F0 contour because this label would imply a change in the rising F0 slope during the rise (i.e., from **L+ !H*** to **!H* +H**). However, as described in detail in the literature on Greek intonation (Arvaniti and Baltazani 2005; Arvaniti, Ladd, and Mennen 1998, 2006), the prenuclear pitch accent in Greek is a simple rising tone, with a low F0 valley anchored at the preaccented syllable and a high f0 peak anchored at the postaccented syllable. Because there is no slope change in the middle of the rising F0, this rising contour would be labeled as **L+ H** in RaP (i.e., an **L+** on the pre-accented syllable and an **H** on the postaccented syllable) without any intervening tone. Consequently, the prenuclear pitch accent in Greek would not be represented as a pitch accent in RaP at all; instead, only the prominence of the accented syllable would be marked (as **x**, for "weak prominence level") on the rhythm tier. In sum, their proposal of L+ !H* +H does not capture the F0 contour of the prenuclear pitch accent in Greek, illustrating that the AM$^+$/RaP system encounters a problem of undergeneration. This problem would disappear if pitch accent were not defined only by pitch "movements" but by a combination of pitch and other acoustic properties such as increased intensity and duration as well as the clarity of segment articulation. This then could also correctly capture the cases where a sequence of three prominent words is realized in the same pitch height.

Finally, RaP is designed to be phonetically transparent and thus make the surface F0 contours more easily recoverable from its labels. Yet the system's emphasis on syntagmatic tonal relationships to the exclusion of all paradigmatic height contrasts may actually hinder recoverability in some cases. For example, consider AM$^+$'s use of the **E** symbol, which refers to a tone having an F0 target equal in height to that of a previous tone. Because **E** only means "equal to previous F0," the symbol **E** itself is not phonetically transparent; we need to look at the following or preceding tone label to know the value of **E**. Thus, **H* +E** means **E** = H, and **E* +L** means **E** = H or !H (i.e., any tone but L), but the interpretation of **E** in a label such as **:E E+ !L*** is not as simple. It would mean a phrase-initial flat F0 followed by a slight lowering of F0. This tells us about the shape of an F0 contour but not about the phonetic value of the F0 plateau, which can be high, mid, or even low. Similarly, because the five tonal symbols in RaP are all representing relative F0 heights in comparison to the previous and sometimes following F0, recovering the F0 values for the symbols would require a much larger utterance domain (or a greater number of symbols) than would be required if using ToBI labels. And the idea that such a wide domain is perceptually necessary is in contradiction to the findings from studies on the perception of pitch targets that Dilley and Breen cite in their commentary (i.e., Bishop and Keating 2012; Honorof and Whalen 2005). It seems unlikely that these problems can be avoided unless tonal labels reflect both paradigmatic and syntagmatic tonal relationships.

Conclusion

Presenting AM$^+$ theory as an "enhanced AM" theory implies that it is a phonological theory. However, a phonological theory of intonation should make predictions about possible or impossible tonal contours of a language. It is not clear how AM$^+$ does this. It is also not clear whether the AM$^+$ model of English intonation can accommodate any fixed inventory of tonal elements or a syntax of their combination. With such a rich set of tonal symbols generated from syntagmatic tonal features and diacritics, AM$^+$ theory would certainly "facilitate" phonetically transparent prosodic annotation. However, as a theory of linguistic pitch, it captures more tonal contrasts than is found in any language, that is, it overgenerates, and in some cases it does not capture distinctive tonal contrasts, that is, it also undergenerates. If a model is only adequate for describing F0 contours by using phonetically transparent labels but that model cannot make predictions about possible or impossible tonal contours, there is no sense in which the model is truly a phonological model. That being the case, it seems that the central difference between AM and AM$^+$ is not the absence versus presence of syntagmatic tonal relationships in phonology—or about the difference in the division of labor between the phonetics and the phonology—but instead whether what is being modeled is even phonology or not.

However, given that RaP is a hybrid transcription system of phonology and phonetics, the issue is not straightforward. Though the "rhythm" aspect of RaP was not described in great detail by Dilley and Breen in their commentary, RaP's rhythm tier encodes the prominence relations among words, as well as the prosodic structure of an utterance—both of which are phonological properties. (These two types of information are labeled on the phonology tier in the PoLaR [points, levels and ranges] annotation system [Ahn, Shattuck-Hufnagel, and Veilleux 2018], which is the most recent annotation system developed to address some of the challenges of the ToBI system.) The pitch tier in RaP, however, captures a mixture of phonological and phonetic properties. In terms of phonological properties, it allows for (i) the transcription of prominence-lending pitch changes (the equivalent to pitch accents in their system, marked with *) and (ii) the association between unstarred tones and starred tones (marked with **+**), distinguishing a bitonal pitch accent from a singleton pitch accent. In terms of phonetic properties, it allows for (i) transcribing small changes in pitch level (marked by **!**) in both upward and downward directions relative to the preceding F0 target and (ii) transcribing any changes (in direction or slope) displayed in an F0 contour. These changes are transcribed without any regard to whether they are contrastive (e.g., labeling phrase-initial medium level F0; labeling H* as an F0 rise, **L+ H**). Therefore, it is important to highlight that RaP is not simply a separation of ToBI's tones tier into two tiers, rhythm and pitch. Although RaP's rhythm tier includes information very similar to the marking of prominence and metrical/prosodic structure on ToBI's tones tier, RaP's pitch tier includes tonal labels that are much more phonetic than ToBI's tonal categories. In fact, the tonal labels on RaP's pitch tier are similar to those of INTSINT, in that both systems label phrase-medial F0 changes based on syntagmatic tonal relationships. At the same time, RaP's pitch tier includes more phonological information about F0 than INTSINT does, because the pitch tier represents an association between tones based on their prominence and alignment to a syllable. However, even though RaP's pitch tier is more phonological than INTSINT, it is less phonological than IPrA because RaP's pitch labels annotate finer differences in F0 than IPrA does.

In sum, ToBI* was designed to be a phonological transcription system, using L and H tonal symbols that capture contrastive tonal events marking either prominence relationships among words or the metrical and prosodic structure of an utterance. As such, the goal of a ToBI transcription is not to recover surface F0 contours like a close copy, but to represent and analyze contours in terms of contrastive tonal categories. Moreover, ToBI is not only a tool for a phonological transcription of an utterance's prosody; it is also a powerful tool for studying prosody and intonation and their interaction with various phenomena in any subfield of linguistics. However, because many researchers in recent years have acknowledged the need to transcribe nondistinctive surface tonal events in addition to the distinctive tonal events, developing transcription systems for nondistinctive F0 events should be encouraged, and these transcription systems should be used in addition to ToBI, but not as a replacement for ToBI transcription.

Note

1. Dilley and Breen say that the phonological representations in AM^{+} are based on two features, [+/– same] and [+/– higher], but later, they introduced [+small] to capture a slope change in rising or falling F0 in English and Greek intonation contours. Yet both [+/–] feature values of [small] are included in their representation of paradigmatic tone specifications, as shown in their table 4c.1.

References

Ahn, B., S. Shattuck-Hufnagel, and N. Veilleux. 2018. "Polar Annotation Conventions: A Tool for Annotating Prosodic Variation." Poster presented at Experimental and Theoretical Advances in Prosody 4, Amherst, MA, October 11–13.

Arvaniti, A., and M. Baltazani. 2005. "Intonational Analysis and Prosodic Annotation of Greek Spoken Corpora." In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun, 84–117. New York: Oxford University Press.

Arvaniti, A., D. Ladd, and I. Mennen. 1998. "Stability of Tonal Alignment: The Case of Greek Prenuclear Accents." *Journal of Phonetics* 26:3–25.

Arvaniti, A., D. Ladd, and I. Mennen. 2006. "Tonal Association and Tonal Alignment: Evidence from Greek Polar Questions and Contrastive Statements." *Language and Speech* 49:421–450. http://dx.doi.org/10.1177/00238309060490040101.

Beckman, M., and G. M. Ayers Elam. 1997. "Guidelines for ToBI Labeling." Unpublished manuscript, Ohio State University. https://www.ling.ohio-state.edu/research/phonetics/E_ToBI/.

Beckman, M., J. Hirschberg, and S. Shattuck-Hufnagel. 2005. "The Original ToBI System and the Evolution of the ToBI Framework." In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun, 9–54. Oxford: Oxford University Press.

Beckman, M., and J. Pierrehumbert. 1986. "Intonational Structure of English and Japanese." *Phonology Yearbook* 3:255–309.

Bishop, J., and P. Keating. 2012. "Perception of Pitch Location within a Speaker's Range: Fundamental Frequency, Voice Quality, and Speaker Sex." *Journal of the Acoustical Society of America* 132 (2): 1100–1112.

Bolinger, D. L. 1951. "Intonation: Levels versus Configurations." *Word* 7 (3): 199–210.

Hirst, D., and A. Di Cristo. 1998. *Intonation Systems: A Survey of Twenty Languages*. Cambridge: Cambridge University Press.

Honorof, D. N., and D. H. Whalen. 2005. "Perception of Pitch Location within a Speaker's F0 Range." *Journal of the Acoustical Society of America* 117 (4): 2193–2200.

Hualde, J. H., and P. Prieto. 2016. "Towards an International Prosodic Alphabet (IPrA)." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7 (1): 1–25.

Pierrehumbert, J. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Pierrehumbert, J., and M. Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Trager, G. L., and H. L. Smith. 1951. *An Outline of English Structure*. Norman, OK: Battenberg Press.

5 Prosody in Articulatory Phonology

Jelena Krivokapić

5.1 Introduction

The linguistic term *prosody* has been conceived of as the structure of the utterance encoding prominence and hierarchical structure at the lexical and phrasal levels (e.g., Beckman 1996; Ladd [1996] 2008, 2001; Shattuck-Hufnagel and Turk 1996; Jun 2005). *Prominence* indicates a prominent syllable in a word or a prominent word in a phrase. The hierarchical structure groups gestures or segments into larger prosodic units, ranging from the smallest (e.g., the syllable in English) to the largest (e.g., the intonation phrase in English). This chapter examines phrase-level prosody as it is understood in the framework of Articulatory Phonology (Browman and Goldstein 1992, 1995; Goldstein, Byrd, and Saltzman 2006) and implemented in the Task Dynamics computational model (Saltzman and Munhall 1989; Saltzman 1995; Nam and Saltzman 2003; Saltzman et al. 2008). This is a relatively recent model of prosodic structure, with its first developments in Byrd et al. (2000). I start with a brief introduction to Articulatory Phonology.[1]

Articulatory Phonology posits that phonological representations and speech production events are isomorphic. The motivation for this approach is to overcome the division that is traditionally made between cognitive and physical aspects of speech (Fowler 1980; Browman and Goldstein 1995) by viewing these two aspects of speech not just as compatible, but as the macroscopic and microscopic aspect of the same representation, where microscopic properties refer to the physical characteristics (articulatory, acoustic) of the macroscopic cognitive units (the combinatorial units of speech). Importantly in this characterization, the microscopic and macroscopic properties of a system are expected to interact with and inform each other (Browman and Goldstein 1990b). The basic unit of speech is a gesture, a linguistically relevant constriction of the vocal tract, which is both an abstract, cognitive unit and at the same time allows for deriving the physical properties of speech. In other words, gestures are simultaneously cognitive units and "units of action" (Browman and Goldstein 1989).

Gestures are modeled as dynamical systems (specifically, constriction gestures are modeled as critically damped mass-spring systems).[2] The articulatory trajectories are lawfully derived by the dynamical systems of the cognitive units (gestures). Another fundamental characteristic of the model is that cognitive processes of speech are understood as dynamical:

> Thus, a dynamical perspective on coordinated movement not only reduces the conceptual distance between cognition on the one hand and mere bodily movement on the other, it forces reconceptualization of the nature of the inner cognitive processes themselves in dynamical terms. It thus turns out that cognition is not best thought of as

> something fundamentally distinct from movements of the body; rather, bodily coordination (and thereby interaction with the world) is really part of cognition itself. (Port and van Gelder 1995, 159)

Each gesture is specified by a set of spatial and temporal parameters (target, stiffness, and damping, where the target corresponds to the linguistic task) and is implemented by the coordinated movements of the articulators associated with each gesture. The *gestural score*, which represents the spatiotemporal properties of gestures in an utterance and their relative timing, drives the movements of the articulators.[3]

It needs to be pointed out that gestures are defined in terms of abstract linguistic tasks (e.g., a lip aperture constriction for [m] rather than the specific movements of lips) and realized through coordinative structures—functional groupings of specific articulators that perform the task (e.g., upper and lower lip and jaw for bilabial constrictions). As cognitive units, gestures are also contrastive, and the contrast can be (i) that gestures can be present or absent, (ii) in terms of their spatiotemporal properties (e.g., gestures using the same articulator can differ in constriction location), and (iii) determined through the temporal organization with other gestures.

Gestures are thus discrete units, specifying a task, but they result in continuous gesture trajectories, which in turn result in continuous movement of the vocal tract articulators. Because gestures are specified for both phonological and phonetic information, there is no need for a component mediating between phonology and phonetics.

Each gesture has associated with it a planning oscillator, or clock, that triggers the activation of that gesture (see Saltzman and Byrd 2000; Nam and Saltzman 2003; Nam, Goldstein, and Saltzman 2009). Thus the timing between two gestures is controlled by the coupling of their planning oscillators (note that the coupling specification is part of the lexical entry).

The model makes use of two prevalent modes of coupling in skilled action, namely in-phase (simultaneous) and anti-phase (sequential) coupling. Both modes are available to humans spontaneously and require no learning, with the in-phase mode being the preferred, more stable mode (Haken, Kelso, and Bunz 1985; Turvey 1990). These two modes have been shown in experimental and modeling work to give rise to syllable structure (Browman and Goldstein 2000; Nam and Saltzman 2003; Goldstein, Byrd, and Saltzman 2006; Nam 2007). When the planning oscillators are in-phase (as in the consonant [C] and vowel [V] gestures in a CV syllable), the associated gestures will be triggered simultaneously, and when the planning oscillators are anti-phase, as in the coda V and C gesture in a CVC syllable, the V and the coda consonant gesture are triggered sequentially.

In complex syllable onsets (but not in codas), for example, in CCV syllables, the planning oscillators for each consonant are in-phase coupled to the vowel and anti-phase coupled to each other (figure 5.1). This leads to competitive coupling, which results in the first C starting earlier and the second starting later (when compared to the CV timing) capturing the *c-center effect*, whereby the midpoint of the onset consonants has a stable temporal relationship to the vowel, regardless of the complexity of the onset sequence (Browman and Goldstein 1988, 2000; Goldstein et al. 2009). In syllable codas, consonants are timed anti-phase to each other, and the first consonant is timed anti-phase to the vowel. Because there is only one type of coupling involved, there is no competitive coupling and consequently no c-center effect.

In the Articulatory Phonology framework, intergestural timing thus emerges from the dynamics of the planning oscillator ensemble, and syllable structure is the result of temporal coordination among the gestures comprising the syllable. The two coupling

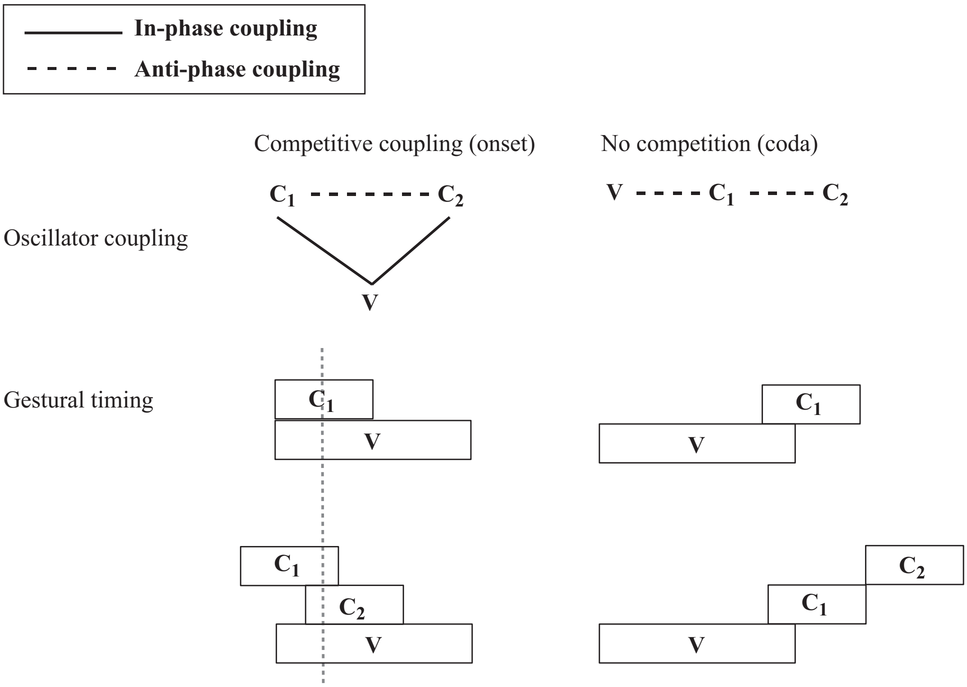


Figure 5.1
Schematic representation of oscillator coupling and the associated gestural timing relationships. Competitive coupling (which exists in syllable onsets) is shown in the left part of the figure, and the gestural timing can be, for example, for the syllable onsets in *pin* and *spin*. The dashed line in the gestural timing of the onsets represents the c-center, which stands in a stable relationship to the vowel regardless of the number of consonants in the onset. The left part of the figure shows coupling in syllable codas, and the examples of words having this gestural timing could be *tip* and *tips*. *Source*: Figure adapted from Marin and Pouplier (2010).

modes have been shown to provide a good account for empirically observed gestural timing in syllables (Browman and Goldstein 1988; Sproat and Fujimura 1993; Krakow 1999; Byrd et al. 2009), although there are cross-linguistic differences and exceptions, and the results are overall less consistent for codas than for onsets (see, e.g., Goldstein et al. 2009; Marin and Pouplier 2010; Hermes 2013; Brunner et al. 2014; and discussion in Pouplier 2011). This conceptualization of the syllable can also help explain a wide range of phenomena, such as typological properties of syllable structure and their acquisition (Goldstein, Byrd, and Saltzman 2006; Nam et al. 2009) and speech errors (Goldstein et al. 2007). (See the overview in Goldstein, Byrd, and Saltzman 2006; Gafos and Goldstein 2011; and Pouplier 2011.)

Finally, it should be mentioned that the specific predictions the model makes about articulation can be tested through the software implementation of the model, TADA (Task Dynamics Application). TADA, which also incorporates the Haskins articulatory synthesizer CASY (Configurable Articulatory Synthesis), takes a text string as the input and generates an intergestural coupling graph (which specifies the gestures of the utterance and the coupling between the planning oscillators of the gestures), the gestural score, the movements of the articulators, and a time-varying vocal tract shape and area function. TADA can be combined with the synthesizer HLsyn (Hanson and Stevens 2002) to synthesize the full acoustic signal. It can also be used without HLsyn, but it will not generate the full acoustic signal of the utterance, only its resonance pattern. For examples of use of TADA, see Gao 2008; Proctor 2009; and Mücke et al. 2012, who used it to verify their predictions for tones in Mandarin, for liquids, and for pitch accent alignment, respectively.[4]

Building on the outlined assumptions and foundations, prosodic structure is understood within Articulatory Phonology as arising through the coordination of a set of prosodic gestures. Like constriction gestures, these gestures present cognitive units and characterize the physical properties of speech. The prosodic gestures are temporal or clock-slowing gestures and tone gestures. We discuss these in sections 5.2 and 5.3, respectively, and prosodic gesture coordination in section 5.4. We turn to polysyllabic shortening and rhythm in section 5.5 and provide a summary and outlook in section 5.6.

5.2 Temporal Modulation Gestures and the π-Gesture Model

Byrd and Saltzman (2003; see also Byrd et al. 2000 for an earlier version of the model) extend the Articulatory Phonology approach to prosodic boundaries and propose that these boundaries are viewed as prosodic gestures (π-gestures). Note that the π-gesture is the cognitive representation of the prosodic boundary, not just its phonetic implementation, in the same spirit that constriction gestures are simultaneously units of representation and units of action. Like constriction gestures, the π-gesture extends in time. It is coordinated with other gestures, and it has a target.[5] Unlike constriction gestures, it does not have a constriction target; instead, it has a temporal target. The goal of the π-gesture is to locally slow the utterance clock that controls the time course of the constriction gestures. The effect of the π-gesture is to slow the time course of the constriction gestures that are active at the same time as the π-gesture. A slower course of activation leads to gestures being temporally longer.[6] Due to the slowdown of the clock, constriction gestures coactive with the π-gesture will also start temporally later compared to the same sequence of gestures without the π-gesture. As a consequence, gestures at boundaries become temporally longer and less overlapped. Both of these effects have been documented in the literature. (For less overlap, see McClean 1973; Hardcastle 1985; Jun 1993; Byrd et al. 2000; Bombien et al. 2010; Bombien, Mooshammer, and Hoole 2013; Byrd and Choi 2010. For lengthening in articulatory studies, see, e.g., Fougeron and Keating 1997; Byrd and Saltzman 1998; Byrd 2000; Keating et al. 2004; Cho 2006; Tabain 2003. For acoustic studies, see, e.g., Oller 1973; Wightman et al. 1992; Shattuck-Hufnagel and Turk 1998.)

The strength of activation of the π-gesture (indicated by the arrow in figure 5.2) determines how strong the boundary effects are. Hierarchically higher boundaries will have stronger activations, and consequently there will be more lengthening and less overlap than at lower boundaries, thus accounting for the empirical findings that boundary lengthening is cumulative. Whether the π-gesture also differs in scope at different boundaries is an empirical question, but the default assumption is that it does not (Byrd and Saltzman 2003; see also Byrd, Krivokapić, and Lee 2006). Note that the π-gesture extends both phrase-finally and phrase-initially; in other words, it accounts for both phrase-final and phrase-initial lengthening.

A potential consequence of the temporal effects are changes in the spatial properties of gestures.[7] Specifically, some of the phrase-initial articulatory effects that have

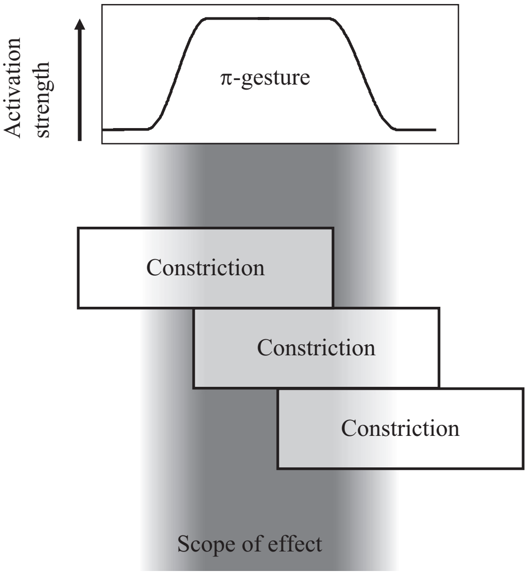


Figure 5.2
A schematic representation of the π-gesture model showing the overlap of the π-gesture and constriction gestures. The arrow indicates the strength of activation of the gesture, and the shading represents the scope of its effect, with darker shading indicating stronger activation.
Source: Figure adapted from Byrd, Krivokapić, and Lee (2006).

been reported in the literature, including for example the increase in gestural amplitude in phrase-initial positions, could be a consequence of the temporal properties (for an overview of initial strengthening phenomena, see Fougeron and Keating 1997; Fougeron 2001; Byrd and Saltzman 2003). One possibility for this effect to arise is when two gestures share the same articulator and are both affected by the π-gesture. Due to the effect of the π-gesture, the two gestures are less overlapped, and there will be more time for the first gesture to reach its target (i.e., there will be detruncation), leading to larger amplitude of that gesture. In that case, the spatial changes can be understood as a consequence of the temporal changes, not as a main effect of prosodic structure. Note, however, that the π-gesture is not meant to capture all the effects of prosodic boundaries, only the temporal effects. For example, while voice onset time (VOT) increases at prosodic boundaries in Korean (Cho and Keating 2001), which could be understood as the result of a π-gesture–induced decrease in gestural overlap (as discussed in Byrd and Saltzman 2003), in Dutch the effect is opposite (VOT decreases at boundaries; Cho and McQueen 2005), indicating that processes other than lengthening take place at prosodic boundaries and interact with the lengthening processes.

We now turn to some of the predictions of the π-gesture model (as developed in Byrd and Saltzman 2003). Like constriction gestures, π-gestures extend over a certain period, and consequently the boundary effects are expected to *extend over a certain period*. The scope of the effect has not been fully established yet (we will return to this question), but the prediction of the model is that the effect will be *local*, because only gestures coactive with the π-gesture will be affected. A related prediction is that there will not be gestures that are “skipped” in the lengthening process. A further prediction of the model is that the boundary effect is strongest at the boundary and will *decrease* *with distance from it*, as the strength of activation of the π-gesture decreases (see figure 5.2). Thus gestures (and acoustic segments) closest to the boundary will lengthen most, and gestures farther away from the boundary will lengthen less, regardless of the specific gestures (and acoustic segments) involved. These predictions have been borne out in a number of articulatory (Byrd, Krivokapić, and Lee 2006; Krivokapić 2007; Katsika 2012 for Greek) and acoustic studies (Berkovits 1993a, 1993b for Hebrew, Cambier-Langeveld 1997 for Dutch).

An effect of the π-gesture is that gestures coactive with it are less overlapped. This is due to the assumption that gestures coactive with a π-gesture have a delayed activation in comparison to phrase-medial gestures. The prediction that arises is that the stronger the boundary, the less the gestural overlap will be. This prediction has been borne out so far (Byrd and Choi 2010; Bombien et al. 2010).

It should be pointed out that the effects of the boundary also depend on the shape of the π-gesture. That is, the strongest effects are predicted to be at the point where the activation of the π-gesture is strongest. In Byrd and Saltzman (2003) it was assumed that the strongest activation of the π-gesture is at the center, between two prosodic phrases. A different shape of the gesture would give different predictions as to where the temporal effects are the strongest. For example, if the shape of the π-gesture is skewed to the left, the effect would be strongest phrase-finally (see discussion and computational simulations in Byrd and Saltzman 2003). At this point, the shape of the π-gesture is an empirical question.

It is expected that any type of gesture coactive with the π-gesture will lengthen, but the possibility cannot be excluded that lengthening could interact with other factors (such as perceptibility or coarticulatory resistance). This prediction has not been extensively investigated, but research so far suggests that all gestures examined lengthen, even though there are some differences between them regarding the extent of lengthening. Thus, Fougeron (2001) finds, for French, postboundary lengthening for a whole range of consonants (/t, n, k, l, s/) and two different vowels; similarly, Cho and McQueen (2005) find final lengthening for /t, d, s, z/ for Dutch. Both velar and bilabial consonants have been found to lengthen as well in German (Bombien, Mooshammer, and Hoole 2013). On the other hand, Fougeron finds less lengthening on /s/ than for other consonants, and Cho and McQueen find that VOT in Dutch is shorter at higher prosodic boundaries than at lower boundaries (the opposite holds for Korean; e.g., see Cho and Keating 2001 and Jun 1993). These might be examples of how lengthening effects interact with other aspects of speech production.

A further prediction of the π-gesture model is that there is no qualitative difference between phrase-final and phrase-initial lengthening. In principle, there might be a difference in the *amount* of lengthening between phrase-final and phrase-initial position. As mentioned, the amount of lengthening depends on the strength of activation of the π-gesture, and therefore, where the effect is strongest (phrase-finally, phrase-initially, or distributed symmetrically) depends on the shape of the π-gesture and its coordination with other gestures (see section 5.4). The shape of the π-gesture is still an open empirical question, but importantly, there is no systematic, categorical, linguistically relevant difference between phrase-final and phrase-initial lengthening. Of course, listeners might be exploiting the lengthening arising from the π-gesture in different manners depending on whether the lengthening occurs at the beginning or at the end of a phrase. For example, phrase-initial lengthening might be more used for word recognition and phrase-final lengthening for syntactic disambiguation. Speakers also might use prosodic boundaries to plan the upcoming utterance, and the planning might be differently distributed across the boundary. But the phonetic and phonological characteristics of the boundary are not specified separately for phrase-initial and phrase-final lengthening.

Within this model, boundaries of different strength also do not differ in a linguistically relevant manner. In other words, there are no different types of π-gestures for different prosodic categories (e.g., for the intermediate or intonation phrase in Beckman and Pierrehumbert's 1986 model); the π-gestures differ only in strength, that is, in the amount of effect. A related issue is whether boundaries can be conceived of as being categorically different, reflecting a prosodic hierarchy with a small number of distinct categories, or whether boundaries are gradient, reflecting a more gradient view of the prosodic hierarchy. The dynamic specification of the π-gesture allows for a structurally gradient or categorical prosodic hierarchy (Byrd 2006). As discussed in Byrd (2006), within Articulatory Phonology, a gradient structure would arise if the π-gesture has a continuum of activation strength values, while a categorical structure could arise by specifying a small number of attractors for the activation strength of the π-gesture. Another way a categorical structure could arise is if the π-gesture has a continuum of activation strength values, but different prosodic categories are distinguished through tone gestures, such as phrase accents and boundary tones (see Byrd 2006; Krivokapić 2007). This is a fundamental question for prosodic theory, but it has not been addressed thoroughly and is in need of extensive research.

Phonetic evidence for a categorical view of prosodic boundaries would include the existence of specific phonetic properties marking each category. For example, the intermediate phrase (ip) and intonation phrase (IP) categories in Beckman and Pierrehumbert's (1986) model are distinguished by the presence of a boundary tone at the IP but not at the ip level (the increase in lengthening between ip and IP does not distinguish them, as this is a gradient, rather than a categorical difference; see also Frota 2000 on this point). Thus, in Beckman and Pierrehumbert's model, phonetic properties give rise to the hierarchy of word, ip, and IP. However, studies examining the production and perception of boundaries indicate that the boundary values form a continuum, not a small number of clusters that would be expected if the hierarchy consists of categorically distinct phrases (Ladd 1988; de Pijper and Sanderman 1994; Swerts 1997; Krivokapić 2007; Krivokapić and Byrd 2012; see also arguments in Ladd [1996] 2008; Frota 2000; and Wagner 2005). Importantly, evidence suggests that the continuum is driven by structural properties of the hierarchy (Ladd 1988; Frota 2000; Krivokapić 2007; Krivokapić and Byrd 2012; Ladd [1996] 2008; Wagner 2005). A second way to establish categorical differences between prosodic categories would be to show that different prosodic categories have different cognitive functions (similarly to phonemic contrasts, which distinguish lexical items). At this point, no evidence of such different functions exists. For example, prosodic phrases are known to disambiguate syntactically ambiguous utterances, but both ip and IP categories do that (e.g., Carlson, Clifton, and Frazier 2001; Clifton, Carlson, and Frazier 2002; Snedeker and Casserly 2010). A third way that a particular category could be justified is by the existence of phonological rules that apply only at that specific level of the prosodic hierarchy. However, claims about such rules have been made based on impressionistic evaluations, rather than on experimental evidence (as pointed out in Byrd 2006 and Wagner 2005). The few studies that have examined these rules have found that they apply in a gradient manner (see discussion in Byrd 2006, Wagner 2005, Krivokapić and Byrd 2012). Thus, questions about the structural properties of the prosodic hierarchy, specifically whether it is gradient or categorical, and if categorical, how many categories exist remain open. Importantly, however, the dynamic view of prosodic boundaries allows us to conceive of a different type of prosodic hierarchy, one that would possibly better fit the empirical evidence to date.

A question related to the structure of the prosodic hierarchy that has not been addressed directly in this model is the relationship between syntactic and prosodic structure. In general, there are two types of approaches to the prosody-syntax interface. In one, phonological and phonetic phenomena associated with the phrase level are mediated by prosodic structure (e.g., Nespor and Vogel [1986] 2007; Selkirk 1984, 2011), while in the other, they are defined directly over syntactic structure, rendering a separate prosodic structure unnecessary (Cooper and Paccia-Cooper 1980; Wagner 2005; see discussion of the approaches in Selkirk 2011). The relationship between syntax and prosody in the Articulatory Phonology model has largely been unexplored, and there is no theoretical reason for the model to prefer one or the other approach. (For a brief exploration of how the mapping between syntactic structure and prosodic structure could be approached within the model, see Krivokapić 2007.)

It has been suggested that the π-gesture model could be extended to account for lengthening in prominence (Byrd and Saltzman 2003; see also Bombien, Mooshammer, and Hoole 2013), and recent work has modeled lexical stress using a clock-slowing gesture, namely the μ-gesture (Saltzman et al. 2008).[8] Before the dynamics of the μ-gesture can be fully developed, a clearer understanding of the gestural timing and the kinematic properties of gestures in prominent syllables is needed. One difference between the μ-gesture and the π-gesture that can be expected lies in the coordination of the temporal modulation gesture to the constriction gestures. The evidence so far indicates that the strongest effect of the boundary is between two phrases and decreases with distance from that point, indicating that the strongest activation of the π-gesture is between two phrases. Prominence on the other hand seems to exert the strongest influence on the vowel of the prominent syllable (Cho and Keating 2009, Bombien et al. 2010; Bombien, Mooshammer, and Hoole 2013), indicating that the strongest activation of the temporal modulation gesture in prominence is at the vowel. This difference does not imply that the temporal modulation gestures differ qualitatively for boundaries and for prominence, just that their coordination with constriction gestures is likely to be different.

To summarize, the π-gesture model, which extends the gestural account of Articulatory Phonology to prosodic boundaries and to prominence, conceptualizes prosodic boundaries in a fundamentally different way from other models. Boundaries are understood to be gestures; they are inherently temporal. That is, their effect is to slow down gestures that are coactive with them; in this way increasing their duration (and in some cases their spatial extent) and reducing gestural overlap. The effect is thus local and continuous, and gradually decreases with the distance from the strongest activation point of the π-gesture. From this model, a set of specific predictions about the temporal properties of prosodic boundaries emerges, and these predictions have been tested in a large number of studies. These studies overall indicate that the gestural approach is a good way to systematically capture properties of boundaries. The model also raises questions about critical aspects of the prosodic hierarchy. A number of questions related to the properties of the π-gesture have been mentioned that still remain to be explored (see Byrd and Saltzman 2003 for additional questions).

5.3 Tone Gestures

Gao (2008; see also Gao 2009) further develops the gestural approach to prosodic structure and extends it to lexical tones. In her approach, tones are gestures that are, like other gestures, specified as dynamical systems (Saltzman 1986; Saltzman and Kelso 1987; Saltzman and Munhall 1989), extend in time, have targets, and are coordinated with other gestures. Gestures generally have specific articulators associated with them, but, as Gao points out, at this point it is not possible to model the articulators involved in the production of F0 due to a lack of understanding of all physiological processes involved in its production (see Hirose 2010 for an overview of the production of F0). However, the *abstract linguistic tasks* of tone gestures can be specified at the level of F0, and in that sense, F0 is an appropriate task variable, just like lip aperture, for example.

Tone gestures have as their goal linguistically relevant variations in fundamental frequency (Gao 2008; McGowan and Saltzman 1995). In modeling the four tones of Mandarin, Gao posits two tone gestures (with a high [H] and with a low [L] target). Tones 1 and 3 have a single gesture (tone 1 has an H target, tone 3 an L target), while tones 2 and 4 arise from a combination of two tonal targets. Tones, like constriction gestures, have planning oscillators associated with them, and the coordination of tones is determined by general coupling mechanisms. In tone 2, the L and H gesture are coordinated in-phase and the L gesture has a shorter activation interval than the H gesture, while in tone 4, the H and L gesture are coordinated anti-phase. For tone 2, the in-phase coordination results in an overlap of the L gesture with the H gesture while both gestures are active. This overlap results in the mid-level F0 in the first part of tone 2, followed by the rise to H. For tone 4, the anti-phase coordination results in the H tone followed by the L tone, that is, the falling contour of tone 4. Gao's computational implementation of these gestural representations in TADA (Nam et al. 2004) showed that these gestural scores yield F0 contours that match contours obtained in speech (Gao 2008, 44; Xu 1997, 67).

Starting from this view of tones, Gao examines tonal alignment in syllables with different syllable structures (CV and CVC) and with different onset consonants. Kinematic data analysis showed two different coordination patterns of the four tones, but they could both be explained by positing that the tone gestures behave like consonant gestures in their coordination patterns, that is, that they are coordinated to vowel and consonant gestures like onset consonants are. For example, for tones 1 and 3, the onset consonant and the vowel are in-phase coupled as are the tone and the vowel gestures, while the consonant and the tone are anti-phase coupled. The multiple coupling specifications give rise to competitive coupling—for example, for onset consonant gestures, and consequently to the c-center effect. Thus, in tones 1 and 3, the order of the gesture onsets is consonant, vowel, tone, indicating that the consonant gesture shifted leftward and the tone gesture rightward from the vowel onset, as in onset consonant clusters. As Gao points out, the specified coupling relations for tones 1 and 3 are like the ones observed for the onset consonants in English *spin* where the two supralaryngeal consonant gestures (tongue tip for /s/ and lip aperture for /p/) are in-phase coupled to the vowel and anti-phase coupled to each other (see figure 5.1). In Gao's analysis then, tones pattern with consonants in how they are coordinated, and they affect the gestural timing of consonant and vowel gestures. Note that the view that tones pattern with consonants fits well with theories of tonogenesis arguing that at least some tones have their historic origin in consonants (e.g., Hombert, Ohala, and Ewan 1979).

In addition to the observed in-phase coordination of vowel and tone gestures, anti-phase coordinations of the tone with the vowel (patterning like coda consonants) have also been suggested. Thus, Hsieh (2011) suggests that tone 3 consists of an L tone coordinated in-phase with the vowel and anti-phase with the onset consonant and an H tone gesture that is coordinated anti-phase with the vowel, accounting in this way for the behavior of tone 3 in different contexts (Gao 2008 discusses tone 3 in sentence-medial positions only, a context also captured by Hsieh 2011; for another account of tone 3, see Yi and Tilsen 2014).

Gao's (2008) approach to tones has been used to investigate the coordination of tones with constriction gestures in pitch accents (this is referred to as "tonal alignment" in the autosegmental-metrical approach to intonation) in Catalan and Standard German (Mücke et al. 2012), and Italian and Viennese German (Niemann et al. 2011). These studies find that differences and similarities between languages arise from the specification of the coordination patterns of tone and vowel gestures; in other words, they are the result of phonological structure. For example, in a kinematic, acoustic, and modeling study examining pitch accent coordination with vowel and consonant gestures in a variety of syllable structures, Mücke et al. (2012) show that in Viennese German, an L and an H tone are coordinated anti-phase to each other and in-phase to the V gesture of the accented syllable. This leads to competitive coupling, which in turn gives rise to the H gesture shifting rightward, and the L gesture shifting leftward (as for onset consonant clusters; see figure 5.1). In Catalan on the other hand, Mücke et al. show that L and H are also coupled anti-phase to each other, but only the H gesture is coupled in-phase with the vowel. The result is that there is no competitive coupling and that the H gesture starts simultaneously with the V gesture. Importantly, as argued in Mücke et al., these results could be described as the tone gesture being aligned to "the left edge of the onset consonant" (in Catalan) as opposed to the "middle of the onset consonant" in German (221). However, an examination of this question that leverages the gestural approach and the coupling oscillator model can account for the observed differences between languages in a precise manner. Similarly, Niemann et al. (2011) argue that German and Italian differ in the gestures comprising rising tones and in the coupling of gestures. Thus, the rising tone in Italian comprises one gesture (with an H target) that is in-phase coordinated with the vowel gesture, but in German, the rising tone is composed of two gestures (with an L and an H target) that are in-phase coordinated to the vowel gesture and anti-phase to each other. This difference in coupling in the two languages leads to different F0 realizations. Niemann et al. also find that pitch accent tone gestures do not affect the lexically specified CV coordination patterns.

Mücke et al. (2012) observe an important distinction between lexical tone gestures and pitch accent tone gestures: tone gestures in lexical tones, but not in pitch accents, affect the lexically specified CV coordination patterns (cf. the effects of tone on coordination reported in Gao 2008 for Mandarin lexical tone and Niemann et al. 2011 for German and Italian pitch accents). They hypothesize that this could be due to the coupling of tone gestures for pitch accents not being specified lexically, but postlexically, and therefore not being fully integrated into the coupling model, as opposed to the coupling of tone gestures in tone languages, where tones are part of the lexical representation and therefore integrated into the coupling model. Thus, whether tone gesture coupling is specified at the lexical or at the postlexical level could reflect and lead to typological differences.[9]

Tonal alignment has been the subject of extensive research (for an overview, see Ladd 2008). So far, consistent alignment between tonal landmarks (such as dips and peaks) and acoustic events has not been found. Lack of consistent alignment could be due to the relevant alignment landmarks being articulatory rather than acoustic in nature (as suggested in D'Imperio et al. 2007 and Gao 2008), but more importantly, the search for alignment points has often focused on surface rather than on structural properties (as discussed in Ladd 2006 and Gao 2008). To account for tonal alignment, Gao (2008) has taken as the starting point of her examination the assumption that the relevant domain of examination of timing is in terms of the structural properties of speech and that tones are gestures (see Ladd 2006 and Prieto and Torreira 2007, who also suggest that gestures in the sense of Articulatory Phonology might provide a better account of alignment).

Much more work remains to be done to understand how the Articulatory Phonology model can account for different languages and the various factors affecting tonal alignment, such as phonological vowel length (e.g., Ladd, Mennen, and Schepman 2000), syllable structure (e.g., Prieto and Torreira 2007), speech rate (e.g., Prieto and Torreira 2007; Hsieh 2011), and upcoming word boundary/prosodic environment (D'Imperio 2001; Prieto, van Santen, and Hirschberg 1995; Silverman and Pierrehumbert 1990). However, the results so far indicate that systematic phonological differences between languages and dialects, exhibited in different tonal composition and different coupling graphs, can account for surface variability in F0 alignment. Viewing tones as gestures within Articulatory Phonology also has the advantage of presenting tonal information in the same manner as constriction and prosodic information, thus allowing for an integrated investigation of prosodic phenomena. The approach has recently been extended to boundary tones in Greek (Katsika 2012; Katsika et al. 2014).

5.4 The Coordination of Prosodic Events

The coordination of gestures is part of the phonological structure in Articulatory Phonology and is specified by the coupled oscillator model. This model has reconceptualized the notion of the syllable and has been shown to account for timing of lexical tones and pitch accents. The π-gesture has not yet been integrated into the coupled oscillator model, but the coordination of the π-gesture to constriction gestures and other prosodic gestures is critical for understanding prosodic structure within Articulatory Phonology, and recent work has been exploring this question.

One of the predictions of the π-gesture model is that lengthening is continuous; that is, it does not skip any segments or gestures. A related question is how the onset of the boundary (specifically, the π-gesture) is determined. This question is of interest for any theory of prosodic structure, but Articulatory Phonology has a consistent way of framing it, namely, in terms of onset of the π-gesture and its coordination with other gestures.

These issues have been examined in a number of recent studies, and a specific focus has been placed on how preboundary prominence interacts with lengthening (e.g., Turk and Shattuck-Hufnagel 2007; Byrd and Riggs 2008; Katsika 2012, 2016; Katsika et al. 2014). In an acoustic study of English, Turk and Shattuck-Hufnagel (2007; see also Rusaw 2013, who replicated these findings) examined the effect of stress on the onset of final lengthening, testing words with final, penultimate, and antepenultimate lexical stress. They found that final lengthening started earlier in the word when the stressed syllable was farther away from the boundary, indicating that the boundary effects shift leftward toward the stressed syllable. However, unstressed syllables between the stressed and the final syllable were not always lengthened or were lengthened less than the stressed syllable. White (2002) also finds that the scope of lengthening extends from the end of the stressed syllable (in antepenultimate, penultimate, or ultimate position) to the boundary, but that the intervening unstressed syllables lengthen as well, contrary to Turk and Shattuck-Hufnagel (2007) and Rusaw (2013). Evidence for an interaction of boundary and prominence also comes from an articulatory magnetometry study by Byrd and Riggs (2008), who find that the onset of final lengthening can be attracted to the prominent syllable (but the results are speaker-specific). Byrd and Riggs suggest two possible accounts: either that the π-gesture shifts toward the prominent syllable, thus keeping the scope of lengthening constant, or that the π-gesture extends toward the prominent syllable, thus increasing the scope of the π-gesture. Crucially, both accounts specifically link the onset of the boundary to another prosodic event (see also Byrd 2006; Krivokapić 2007), capturing the interdependence of boundaries and prominence that many studies have shown.

For Greek, Katsika (2012, 2016) found in an articulatory magnetometer study that the onset of final lengthening occurred during the final syllable in words with final stress, but that it started earlier when the stressed syllable occurred earlier (consistent with White 2002; Turk and Shattuck-Hufnagel 2007; Byrd and Riggs 2008; and Rusaw 2013). Like White (2002), she found that lengthening is continuous (once it starts, it does not skip any gestures), as predicted by the π-gesture model. Katsika (2012) and Katsika et al. (2014) combined this examination of temporal properties of prosodic boundaries with the examination of the onset of the boundary tone in IP boundaries. The onset of the boundary tone was always during the final vowel, but it occurred earlier within the vowel when the stressed syllable occurred earlier; in other words, it shifted toward the stressed syllable.

Katsika and Katsika et al. suggest that this interdependency of prominence and boundary arises through the coordination of the μ-gesture, the boundary tone gesture, and the π-gesture. The π-gesture has a dual coordination: it is coordinated both with the phrase-final vowel gesture and with the μ-gesture (the temporal modulation gesture on the stressed syllable) of the phrase-final word. The coordination with the μ-gesture is weaker, thus affecting the onset of the π-gesture less than the π-gesture's coordination with the phrase-final vowel gesture. In this way, the onset of lengthening will depend on the position of the stressed syllable; that is, the onset will shift slightly, in a gradient manner, toward the stressed syllable. Thus, the flexible scope of lengthening is accounted for by these two coordinations.[10]

Katsika and Katsika et al. also suggest that the variability in the alignment of the boundary tone can be accounted for if it is assumed that the boundary tone gesture is activated when the π-gesture reaches a certain high level of activation (see also Byrd 2006 on this relationship between π-gesture strength of activation and boundary tone). When lexical stress of the phrase-final word is earlier in the word, the π-gesture will start earlier, and it will reach the level of activation that triggers the boundary tone earlier, thus triggering the boundary tone earlier. This means that, like the π-gesture, the onset of the boundary tone will shift slightly toward the stressed syllable (for further discussion, see Katsika 2012 and Katsika et al. 2014). Thus, the variability in the onset of lengthening and onset of the boundary tone gesture arises from the coordination of the π-gesture and the μ-gesture. This account also indicates a further interdependency of boundary properties: that between the temporal and tonal properties. Note that the coordination of prosodic gestures (which gestures are coordinated, what the strength of the coordination is, what type of coordination) can be expected to be a source of cross-linguistic variation (see Katsika et al. 2014).

Katsika and Katsika et al. further identify pause postures, which are specific configurations of the vocal tract occurring at IP boundaries. They argue that pause postures are triggered by a very high activation of the π-gesture (higher than needed for the activation of the boundary tone gesture, thus achieving that only very strong boundaries have pause postures). While pause postures need to be further examined, some corroborating evidence for pauses as linguistic entities comes from the studies of Ramarayanan et al. (2009, 2010, 2013). These studies show that grammatical pauses, defined there as pauses between syntactic constituents, have specific kinematic properties: they exhibit less variability and a decrease in velocity of articulators in comparison to other types of silent intervals. The observation of these pause-specific configurations also indicates the need for further exploration of the relationship between final lengthening, pauses, and initial lengthening (see also discussions in Byrd and Saltzman 2003; Krivokapić 2014; and Katsika et al. 2014).

To summarize, a number of recent studies (Byrd 2006, Byrd and Riggs 2008, Katsika 2012, Katsika et al. 2014) argue that prosodic boundaries arise through a set of temporally specified prosodic gestures that interact in a dynamic way. This view allows for a characterization of the tonal and temporal properties of prosodic structure, including an account of the gradient effects that prosodic structure gives rise to. Recent work on articulation during pauses has revealed further aspects of prosodic boundaries.

5.5 Polysyllabic Shortening, Rhythm, and the Prosodic Hierarchy

A recent extension of the theory is the modeling of polysyllabic shortening and rhythm. Early researchers (Pike 1945; Abercrombie 1967) developed the notions of stress timing and syllable timing and proposed that all languages are stress-timed or syllable-timed. The phonetic correlates of these two rhythm types were suggested to be foot and syllable isochrony. Since then, a large body of research has examined rhythmic properties of languages but has not found evidence for isochrony in either production or perception (Bolinger 1965; Huggins 1972; Lehiste 1972, 1977; Nakatani, O'Connor, and Aston 1981), and it is now abundantly clear that strict isochrony does not exist (see the overview in Arvaniti 2012).

Despite the clear evidence against strict isochrony, it is not disputed that there are temporal regularities, as evidenced in durational aspects of speech, although it is likely that other aspects of speech, such as the timing of pitch accents and phrasal boundaries, are also relevant for characterizing rhythm (as argued in Jun 2014). Specifically, there is evidence of polysyllabic shortening—the shortening of stressed syllables with the addition of unstressed syllables to a word or a foot—that can be seen as a tendency toward foot isochrony (Lehiste 1972; Klatt 1973; Port 1981; Rakerd, Sennett, and Fowler 1987; Kim and Cole 2005; White and Turk 2010; Shattuck-Hufnagel and Turk 2011). Kim and Cole (2005), for example, find that the duration of the foot (defined in the Abercrombian sense of extending from one stressed syllable to the next) increases with the addition of unstressed syllables. In the process, the unstressed syllables do not shorten, but the stressed syllable does. In this way, the duration of the foot is somewhat constrained.

Within Articulatory Phonology, rhythmic properties are understood to emerge through the coupling of the foot oscillator, the syllable oscillator, and the temporal modulation gesture μ (Saltzman et al. 2008; Nam et al. 2008). The μ-gesture, which is associated with stressed syllables, slows the gestural activation of coactive gestures. As a consequence, stressed syllables are longer than unstressed syllables. The foot and the syllable oscillator (and, similarly, phrase-level oscillators, although this part of the model is not fully developed yet) extend the model of planning oscillators for gestures to prosodic structure.

As Saltzman et al. (2008, 180) stated, “the foot oscillator attempts to keep the duration of the foot constant, while the syllable oscillator attempts to keep the duration of the syllable constant.” These opposite goals of the oscillators, and the resulting competition, are modeled with interoscillator coupling functions, where the foot and syllable oscillators are coupled to each other bidirectionally, and each oscillator has a weighting that specifies the influence of the oscillator in the function. Depending on which oscillator’s weighting is stronger, the foot or the syllable will be the more dominant oscillator, and the language will display a tendency toward foot or syllable isochrony. In the case of foot dominance, an increase in the number of syllables in the foot will lead to the shortening of syllables (the syllables are “squeezed” in the foot), leading to polysyllabic shortening (the approach discussed so far is based on O’Dell and Nieminen 1999; for related approaches see also Barbosa 2002; Cummins and Port 1998; and Tilsen 2009). To account for the fact that only the stressed (that is, not the unstressed) syllable shortens (as found in Kim and Cole 2005), the foot oscillator’s strength (weight) in the interoscillator coupling function is modeled to be weaker during the unstressed than during the stressed syllables. Thus, the syllable oscillator is dominant during the unstressed syllables. This leads to the foot oscillator’s squeezing being weaker during the unstressed than during the stressed syllables, which results in shortening of the stressed but not the unstressed syllables.

The model of Saltzman et al. (2008; see also Nam et al. 2008 and O’Dell and Nieminen 1999) does not predict strict isochrony, but it captures polysyllabic shortening and the tendencies toward isochrony, thus allowing both the empirical facts and the hard-to-capture division between stress-timed and syllable-timed languages to be accounted for. Typological differences between languages can be understood as the difference in the dominance of the foot and syllable oscillator (compatible with a gradiently varying scale of stress-timedness, as Dauer 1983 suggested).

Extending the planning oscillators to prosodic structure provides the possibility of understanding the prosodic hierarchy as a network of oscillators, where oscillators with higher frequencies (for lower prosodic phrases) are embedded into oscillators of lower frequencies (higher prosodic phrases), as discussed in Saltzman et al. (2008; see also Cummins and Port 1998 and Goldstein 2012). This approach has not been fully implemented within the Task Dynamics model of Articulatory Phonology (but see Tilsen’s 2009 dynamical model for this type of approach). The integration of this hierarchy of oscillators with other aspects of prosodic structure (such as temporal and tonal boundary events) is also a matter for future research. One immediately obvious consequence of this conceptualization of the prosodic hierarchy is that as in the π-gesture approach, prosodic phrases based on the planning oscillators do not differ qualitatively; corresponding to the evidence outlined in section 5.1, they differ only in oscillator frequencies.

Note that two possible views of prosodic structure have been outlined: the prosodic hierarchy can be understood as arising from a hierarchy of nested planning oscillators or through the coordination of the π-gesture (which can have a categorical or a gradient activation), other prosodic gestures, and constriction gestures. These approaches also entail two distinct views of prosodic boundaries. In one view (the π-gesture view), the boundary is a phonological unit, and in the other, the boundary is a consequence of phrases, without a separate phonological status. Future research will determine which of these directions represents a more accurate way of conceptualizing the prosodic hierarchy.

5.6 Conclusion

Within Articulatory Phonology, prosodic structure is understood in a gestural manner. Prosodic boundaries are viewed as arising through the coordination of temporal, tonal, and constriction gestures. Rhythmic structure is determined by the μ-gesture and the coupling between foot and syllable oscillators.

Within this model, and as opposed to other prosodic models, temporal information is part of the linguistic representation, and the coordination of gestures is guided by general principles of coupling. The clear, predictive nature of this model has sparked numerous studies. In particular, it has allowed the question of the scope of effect of the boundary and of gestural coordination to be formulated precisely and has led to a number of studies examining the temporal properties of boundaries and the relationship between temporal properties of boundaries, tonal gestures, and prominence. The dynamical nature of the model also makes it possible to account for both gradient and categorical prosodic phenomena (as shown, e.g., in the examination of boundary strength or in the properties of the foot) in a manner that would be difficult to capture in other models. Furthermore, the simultaneous examination of temporal and tonal prosodic events is crucial for any model of prosodic theory, but is facilitated in Articulatory Phonology by the understanding that all prosodic phenomena are gestures that can be coordinated with each other, and in that way can influence one another. This view also suggests an extension of the model to integrate body gestures. Recent studies have found strong evidence that body gestures are timed to prosodic structure (Krahmer and Swerts 2007; Esteve-Gibert and Prieto 2013; Rochet-Capellan et al. 2008). Speech production models, and models of prosodic structure in particular, need to uncover the principles of this coordination. Within the Articulatory Phonology model, body gestures could be understood as gestures—for example, a hand or head gesture. These, like constriction gestures, would be associated with planning oscillators and coordinated with constriction and prosodic gestures following the same principles of coordination as other gestures.

Many aspects of the model remain to be developed, among them many that are not well understood regardless of the theoretical framework. However, the basic assumption of the model—the dynamic, gestural approach in which prosodic units have temporal properties and are coordinated with each other by general coupling mechanisms—should continue to lead to fruitful research.

Acknowledgments

I would like to thank Louis Goldstein, Argyro Katsika, Hosung Nam, and especially Elliot Saltzman for many productive discussions over the past year about questions presented here; Harim Kwon and Kate Sherwood whose questions in my Prosody 2014 seminar provided guidelines in writing this chapter; and Pam Beddor, Susanne Fuchs, and Caterina Petrone for very helpful comments on an earlier version of the manuscript. This work was supported by National Institutes of Health grant DC002717 to Douglas Whalen.

Notes

1. For overviews of the model, see Browman and Goldstein (1989, 1990a, 1990b, 1991, 1992); Goldstein, Byrd, and Saltzman (2006); Pouplier (2011); Gafos and Goldstein (2011); and the

Haskins website: http://www.haskins.yale.edu/research/gestural.html. For an introduction to Task Dynamics, see Hawkins (1992), and for an overview of the application of dynamical systems theory to speech, see van Lieshout (2004).

2. A dynamical system is one in which future states of the system are determined by its present state and a set of laws that specify the forces changing the system (Saltzman and Kelso 1987; Saltzman 1995). The computational model is named "Task Dynamics" to capture that "(a) it deals with the performance of well-learned skilled movements or gestures designed to accomplish real-world tasks; and (b) it is defined with respect to the dynamics that underlie a given action's kinematics. Note that kinematics refers to a gesture's observable spatiotemporal properties (e.g., its position, velocity, and acceleration trajectories over time), while dynamics refers to the pattern of the underlying field of forces that gives rise to these kinematics" (Saltzman 1986, 130).

3. This view of phonology and phonetics does not lend itself to easy transcription, and this applies to the model of prosody as well.

4. TADA can be downloaded at http://www.haskins.yale.edu/tada_download/index.php and the manual for the model (Nam and Goldstein 2007) is available at http://www.haskins.yale.edu/tada_download/doc/TADA_manual_v09.pdf.

5. The π-gesture has not been integrated in the coupled oscillator model at this point, but the coordination of the π-gesture to constriction gestures is a central question of the theory, and I discuss approaches to this question in section 5.4.

6. Gestures also become slower (with lower peak velocities), but this interacts with gesture displacement (or amplitude), which means that the change in peak velocity might not always be observable. See Byrd and Saltzman (2003).

7. A spatial-modulation gesture, which modifies the spatial extents of constriction gestures, is discussed in Saltzman et al. (2008). At this point, however, it has not been used extensively in the model because it is not yet clear whether a separate spatial-modulation gesture is necessary to alter spatial extent, or whether such effects can be derived indirectly via the truncation/detruncation effects that are direct outcomes of applying a temporal modulation gesture (E. Saltzman, personal communication, 2015; see also Byrd and Saltzman 2003).

8. The π-gesture and the μ-gesture are conceptually similar, but the π-gesture has not yet been implemented in the coupled planning oscillator network. Both locally modulate the temporal properties of an utterance. For a further discussion of the entailed differences, see Saltzman et al. (2008). For lengthening under prominence, see, for example, Turk and Sawusch (1997), Turk and White (1999), Cho and Keating (2009), Mücke and Grice (2014), and the overview in Fletcher (2010).

9. While the typological predictions of the prosodic model in Articulatory Phonology have not been much discussed, the coordination between gestures can be expected to be a major source of typological variation. See discussions about typology of syllable structure, and how it relates to coupling graphs, in Nam, Goldstein, and Saltzman (2009).

10. For a related approach see Rusaw (2013), who modeled the interaction between boundary and prominence lengthening, as found in Turk and Shattuck-Hufnagel (2007) and replicated in Rusaw (2013), with an artificial central pattern generator neural network model.

References

Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Arvaniti, A. 2012. "The Usefulness of Metrics in the Quantification of Speech Rhythm." *Journal of Phonetics* 40:351–373.

Barbosa, P. A. 2002. “Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production.” In *Proceedings of the first International Conference on Speech Prosody*, Aix-en-Provence, France. https://www.isca-speech.org/archive/sp2002/papers/sp02_163.pdf.

Beckman, M. E. 1996. “The Parsing of Prosody.” *Language and Cognitive Processes* 11:17–67.

Beckman, M. E., and J. B. Pierrehumbert. 1986. “Intonational Structure in Japanese and English.” *Phonology Yearbook* 3:255–309.

Berkovits, R. 1993a. “Progressive Utterance: Final Lengthening in Syllables with Final Fricatives.” *Language and Speech* 36:89–98.

Berkovits, R. 1993b. “Utterance-Final Lengthening and the Duration of Final-Stop Closures.” *Journal of Phonetics* 21:479–489.

Bolinger, D. L. 1965. “Pitch Accent and Sentence Rhythm.” In *Forms of English: Accent, Morpheme, Order*, edited by I. Abe and T. Kanekiyo, 139–180. Cambridge, MA: Harvard University Press.

Bombien, L., C. Mooshammer, and P. Hoole. 2013. “Articulatory Coordination in Word-Initial Clusters of German.” *Journal of Phonetics* 41:546–561.

Bombien, L., C. Mooshammer, P. Hoole, and B. Kühnert. 2010. “Prosodic and Segmental Effects on EPG Contact Patterns of Word-Initial German Clusters.” *Journal of Phonetics* 38:388–403.

Browman, C. P., and L. Goldstein. 1988. “Some Notes on Syllable Structure in Articulatory Phonology.” *Phonetica* 45:140–155.

Browman, C. P., and L. Goldstein. 1989. “Articulatory Gestures as Phonological Units.” *Phonology* 6:201–251.

Browman, C. P., and L. Goldstein. 1990a. “Gestural Specification Using Dynamically-Defined Articulatory Structures.” *Journal of Phonetics* 18:299–320.

Browman, C. P., and L. Goldstein. 1990b. “Representation and Reality: Physical Systems and Phonological Structure.” *Journal of Phonetics* 18:411–424.

Browman, C. P., and L. Goldstein. 1991. “Tiers in Articulatory Phonology, with Some Implications for Casual Speech.” In *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, edited by J. Kingston and M. E. Beckman, 341–376. Cambridge: Cambridge University Press.

Browman, C. P., and L. Goldstein. 1992. “Articulatory Phonology: An Overview.” *Phonetica* 49:155–180.

Browman, C. P., and L. Goldstein. 1995. “Dynamics and Articulatory Phonology.” In *Mind as Motion*, edited by R. F. Port and T. van Gelder, 175–193. Cambridge, MA: MIT Press.

Browman, C., and L. Goldstein. 2000. “Competing Constraints on Intergestural Coordination and Self-Organization of Phonological Structures.” *Bulletin de la Communication Parlée* 5:25–34.

Brunner, J., Geng, C., Sotiropoulou, S., and Gafos, A. 2014. “Timing of German Onset and Word Boundary Clusters.” *Laboratory Phonology* 5:403–454.

Byrd, D. 2000. “Articulatory Vowel Lengthening and Coordination at Phrasal Junctures.” *Phonetica* 57:3–16.

Byrd, D. 2006. “Relating Prosody and Dynamic Events: Commentary on the Papers by Cho, Navas, and Smiljanić.” In *Laboratory Phonology 8: Varieties of Phonological Competence*, edited by L. Goldstein, 549–561. Berlin: Mouton de Gruyter.

Byrd, D., and S. Choi. 2010. “At the Juncture of Prosody, Phonology, and Phonetics: The Interaction of Phrasal and Syllable Structure in Shaping the Timing of Consonant Gestures.” In *Papers in Laboratory Phonology 10*, edited by C. Fougeron, B. Kühnert, M. Imperio, and N. Vallee, 31–59. Berlin: Mouton de Gruyter.

Byrd, D., A. Kaun, S. Narayanan, and E. Saltzman. 2000. “Phrasal Signatures in Articulation.” In *Papers in Laboratory Phonology V*, edited by M. B. Broe and J. P. Pierrehumbert, 70–87. Cambridge: Cambridge University Press.

Byrd, D., J. Krivokapić, and S. Lee. 2006. “How Far, How Long: On the Temporal Scope of Phrase Boundary Effects.” *Journal of the Acoustical Society of America* 120:1589–1599.

Byrd, D., and D. Riggs. 2008. “Locality Interactions with Prominence in Determining the Scope of Phrasal Lengthening.” *Journal of the International Phonetic Association* 38:187–202.

Byrd, D., and E. Saltzman. 1998. Intragestural Dynamics of Multiple Phrasal Boundaries.” *Journal of Phonetics* 26:173–199.

Byrd, D., and E. Saltzman. 2003. “The Elastic Phrase: Dynamics of Boundary-Adjacent Lengthening.” *Journal of Phonetics* 31:149–180.

Byrd, D., S. Tobin, E. Bresch, and S. Narayanan. 2009. “Timing Effects of Syllable Structure and Stress on Nasals: A Real-Time MRI Examination.” *Journal of Phonetics* 37:97–110.

Cambier-Langeveld, T. 1997. “The Domain of Final Lengthening in the Production of Dutch.” In *Linguistics in the Netherlands*, edited by J. Coerts and H. de Hoop, 13–24. Amsterdam: John Benjamins.

Carlson, K., C. Clifton Jr., and L. Frazier. 2001. “Prosodic Boundaries in Adjunct Attachment.” *Journal of Memory and Language* 45:58–81.

Cho, T. 2006. “Manifestation of Prosodic Structure in Articulation: Evidence from Lip Kinematics in English.” In *Laboratory Phonology 8: Varieties of Phonological Competence*, edited by L. Goldstein, 519–548. New York: Walter de Gruyter.

Cho, T., and P. Keating. 2001. “Articulatory and Acoustic Studies of Domain-Initial Strengthening in Korean.” *Journal of Phonetics* 29:155–190.

Cho, T., and P. Keating. 2009. “Effects of Initial Position versus Prominence in English.” *Journal of Phonetics* 37:466–485.

Cho, T., and J. McQueen. 2005. “Prosodic Influences on Consonant Production in Dutch: Effects of Prosodic Boundaries, Phrasal Accent, and Lexical Stress.” *Journal of Phonetics* 33: 121–157.

Clifton, C. J., K. Carlson, and L. Frazier. 2002. “Informative Prosodic Boundaries.” *Language and Speech* 45:87–114.

Cooper, W. E., and J. Paccia-Cooper. 1980. *Syntax and Speech*. Cambridge, MA: Harvard University Press.

Cummins, F., and R. Port. 1998. “Rhythmic Constraints on Stress Timing in English.” *Journal of Phonetics* 26:145–171.

Dauer, R. M. 1983. “Stress-Timing and Syllable-Timing Reanalyzed.” *Journal of Phonetics* 11:51–69.

de Pijper, J. R., and A. A. Sanderman. 1994. “On the Perceptual Strength of Prosodic Boundaries and Its Relation to Suprasegmental Cues.” *Journal of the Acoustical Society of America* 96:2037–2047.

D'Imperio, M. 2001. "Focus and Tonal Structure in Neapolitan Italian." *Speech Communication* 33:339–256.

D'Imperio, M., R. Espesser, H. Loevenbruck, C. Menezes, N. Nguyen, and P. Welby. 2007. "Are Tones Aligned with Articulatory Events? Evidence from Italian and French." In *Papers in Laboratory Phonology 9*, edited by J. Cole and J. I. Hualde, 577–608. Berlin: Mouton de Gruyter.

Esteve-Gibert, N., and P. Prieto. 2013. "Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements." *Journal of Speech, Language, and Hearing Research* 56:850–864.

Fletcher, J. 2010. "The Prosody of Speech: Timing and Rhythm." In *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle, J. Laver, and F. E. Gibbon, 523–602. Hoboken, NJ: Wiley-Blackwell.

Frota, S. 2000. *Prosody and Focus in European Portuguese: Phonological Phrasing and Intonation*. New York: Garland.

Fougeron, C. 2001. "Domain-Initial Articulatory Variations in French." *Journal of Phonetics* 29:109–135.

Fougeron, C., and P. Keating. 1997. "Articulatory Strengthening at Edges of Prosodic Domains." *Journal of the Acoustical Society of America* 101:3728–3740.

Fowler, C. A. 1980. "Coarticulation and Theories of Extrinsic Timing." *Journal of Phonetics* 8:113–133.

Gafos, A., and L. Goldstein. 2011. "Articulatory Representation and Organization." In *The Handbook of Laboratory Phonology*, edited by A. C. Cohn, C. Fougeron, and M. K. Huffman, 220–231. Oxford: Oxford University Press.

Gao, M. 2008. "Mandarin Tones: An Articulatory Phonology Account." PhD diss., Yale University.

Gao, M. 2009. "Gestural Coordination among Vowel, Consonant and Tone Gestures in Mandarin Chinese." *Chinese Journal of Phonetics* 2:43.

Goldstein, L. 2012. "Dynamical Models of the Prosodic Hierarchy." Paper presented at the Workshop on Prosody of the Wider World, Cologne, Germany, May 30–June 1.

Goldstein, L., D. Byrd, and E. Saltzman, E. 2006. "The Role of Vocal Tract Gestural Action Units in Understanding the Evolution of Phonology." In *Action to Language via the Mirror Neuron System*, edited by M. Arbib, 215–249. Cambridge: Cambridge University Press.

Goldstein, L., H. Nam, E. Saltzman, and I. Chitoran. 2009. "Coupled Oscillator Planning Model of Speech Timing and Syllable Structure." In *Frontiers in Phonetics and Speech Science*, edited by G. Fant, H. Fujisaki, and J. Shen, 239–250. Beijing: Commercial Press.

Goldstein, L., M. Pouplier, L. Chen, E. Saltzman, and D. Byrd. 2007. "Dynamic Action Units Slip in Speech Production Errors." *Cognition* 103:386–412.

Haken, H., J. A. S. Kelso, and H. Bunz. 1985. "A Theoretical Model of Phase Transitions in Human Hand Movements." *Biological Cybernetics* 51:347–356.

Hanson, H., and K. N. Stevens. 2002. "A Quasi-Articulatory Approach to Controlling Acoustic Source Parameters in a Klatt-Type Formant Synthesizer Using HLSyn." *Journal of the Acoustical Society of America* 112:1158–1182.

Hardcastle, W. J. 1985. "Some Phonetic and Syntactic Constraints on Lingual Coarticulation during /kl/ Sequences." *Speech Communication* 4:247–263.

Hawkins, S. 1992. “An Introduction to Task Dynamics.” In *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, edited by G. J. Docherty and D. R. Ladd, 9–25. Cambridge: Cambridge University Press.

Hermes, A. 2013. “Articulatory Coordination and Syllable Structure in Italian.” In *Speech Production and Perception*, edited by S. Fuchs and P. Perrier, 9–25. Frankfurt am Main: Peter Lang Verlag.

Hirose, H. 2010. “Investigating the Physiology of Laryngeal Structures.” In *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle, J. Laver, and F. E. Gibbon, 130–152. Hoboken, NJ: Wiley-Blackwell.

Hombert, J.-M., J. J. Ohala, and W. G. Ewan. 1979. “Phonetic Explanations for the Development of Tones.” *Language* 55:37–58.

Hsieh, F.-Y. 2011. “A Gestural Account of Mandarin Tone 3 Variation.” In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, 890–893. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/index.htm.

Huggins, A. W. F. 1972. “On the Perception of Temporal Phenomena in Speech.” *Journal of the Acoustical Society of America* 51:1279–1290.

Jun, S.-A. 1993. “The Phonetics and Phonology of Korean Prosody.” PhD diss., Ohio State University.

Jun, S.-A. 2005. “Prosodic Typology.” In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun, 430–458. Oxford: Oxford University Press.

Jun, S.-A. 2014. “Prosodic Typology: By Prominence Type, Word Prosody, and Macro-rhythm.” In *Prosodic Typology II: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun, 520–540. Oxford: Oxford University Press.

Katsika, A. 2012. “Coordination of Prosodic Gestures at Boundaries in Greek.” PhD diss., Yale University.

Katsika. A. 2016. “The Role of Prominence in Determining the Scope of Boundary-Related Lengthening.” *Journal of Phonetics* 55:149–181.

Katsika, A., J. Krivokapić, C. Mooshammer, M. Tiede, and L. M. Goldstein. 2014. “The Coordination of Boundary Tones and Their Interaction with Prominence.” *Journal of Phonetics* 44:62–82.

Keating, P., T. Cho, C. Fougeron, and C. Hsu. 2004. “Domain-Initial Articulatory Strengthening in Four Languages.” In *Phonetic Interpretation*, edited by J. Local, R. Ogden, and R. Temple, 143–161. Cambridge: Cambridge University Press.

Kim, H., and J. Cole. 2005. “The Stress Foot as a Unit of Planned Timing: Evidence from Shortening in the Prosodic Phrase.” In *Proceedings of Interspeech 2005*, 2365–2368. https://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_2365.pdf.

Klatt, D. H. 1973. “Interaction between Two Factors That Influence Vowel Duration.” *Journal of the Acoustical Society of America* 54:1102–1104.

Krahmer, E., and M. Swerts. 2007. “The Effects of Visual Beats on Prosodic Prominence: Acoustic Analyses, Auditory Perception, and Visual Perception.” *Journal of Memory and Language* 57 (3): 396–414.

Krakow, R. A. 1999. “Physiological Organization of Syllables: A Review.” *Journal of Phonetics* 27:23–54.

Krivokapić, J. 2007. “The Planning, Production, and Perception of Prosodic Structure.” PhD diss., University of Southern California.

Krivokapić, J. 2014. "Gestural Coordination at Prosodic Boundaries and Its Role for Prosodic Structure and Speech Planning Processes." *Communicative Rhythms in Brain and Behaviour. Philosophical Transactions of the Royal Society B (Biology)* 369:20130397.

Krivokapić, J., and D. Byrd. 2012. "Prosodic Boundary Strength: An Articulatory and Perceptual Study." *Journal of Phonetics* 40:430–442.

Ladd, D. R. 1988. "Declination 'Reset' and the Hierarchical Organization of Utterances." *Journal of the Acoustical Society of America* 84:530–544.

Ladd, D. R. (1996) 2008. *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. R. 2001. "Intonation." In *Language Typology and Language Universals: An International Handbook*, edited by M. Haspelmath, E. König, W. Oesterreicher, and W. Raible, 1380–1390. Berlin: Mouton de Gruyter.

Ladd, D. R. 2006. "Segmental Anchoring of Pitch Movements: Autosegmental Association or Gestural Coordination?" *Italian Journal of Linguistics* 18 (1): 19–38.

Ladd, D. R., I. Mennen, and A. Schepman. 2000. "Phonological Conditioning of Peak Alignment of Rising Pitch Accents in Dutch." *Journal of the Acoustical Society of America* 107:2685–2696.

Lehiste, I. 1972. "The Timing of Utterances and Linguistic Boundaries." *Journal of the Acoustical Society of America* 51:2018–2024.

Lehiste, I. 1977. "Isochrony Reconsidered." *Journal of Phonetics* 5:253–263.

Marin, S., and M. Pouplier. 2010. "Temporal Organization of Complex Onsets and Codas in American English: Testing the Predictions of a Gestural Coupling Model." *Motor Control* 14:380–407.

McClean, M. 1973. "Forward Coarticulation of Velar Movement at Marked Junctural Boundaries." *Journal of Speech, Language and Hearing Research* 16:286–296.

McGowan, R. S., and E. Saltzman. 1995. "Incorporating Aerodynamic and Laryngeal Components into Task Dynamics." *Journal of Phonetics* 23:255–269.

Mücke, D., and M. Grice. 2014. "The Effect of Focus Marking on Supralaryngeal Articulation: Is It Mediated by Accentuation?" *Journal of Phonetics* 44:47–61.

Mücke, D., H. Nam, A. Hermes, and L. Goldstein. 2012. "Coupling of Tone and Constriction Gestures in Pitch Accents." In *Consonant Clusters and Structural Complexity*, edited by P. Hoole, M. Pouplier, L. Bombien, C. Mooshammer, and B. Kühnert, 205–230. Berlin: Mouton de Gruyter.

Nakatani, L. H., K. O'Connor, and C. H. Aston. 1981. "Prosodic Aspects of American English Speech Rhythm." *Phonetica* 38:84–106.

Nam, H. 2007. "Syllable-Level Intergestural Timing Model: Split-Gesture Dynamics Focusing on Positional Asymmetry and Moraic Structure." In *Laboratory Phonology 9*, edited by J. Cole and J. I. Hualde, 483–506. Berlin: Walter de Gruyter.

Nam, H., and L. Goldstein. 2007. *TADA (TAsk Dynamics Application) Manual*. http://www.haskins.yale.edu/tada_download/doc/TADA_manual_v09.pdf.

Nam, H., L. Goldstein, and E. Saltzman, E. 2009. "Self-Organization of Syllable Structure: A Coupled Oscillator Model." In *Approaches to Phonological Complexity*, edited by F. Pellegrino, E. Marisco, and I. Chitoran, 299–328. Berlin: Mouton de Gruyter.

Nam, H., L. Goldstein, E. Saltzman, and S. Byrd. 2004. "TADA: An Enhanced, Portable Task Dynamics Model in MATLAB." *Journal of the Acoustical Society of America* 115:2430.

Nam, H., and E. Saltzman, E. 2003. "A Competitive, Coupled Oscillator of Syllable Structure." Paper presented at the Fifteenth International Congress of Phonetic, Barcelona, August 3.

Nam, H., E. Saltzman, J. Krivokapić, and L. Goldstein. 2008. "Modeling the Durational Difference of Stressed vs. Unstressed Syllables." In *Proceedings of the Eighth Phonetic Conference of China*. https://sites.lsa.umich.edu/jelenakrivokapic/wp-content/uploads/sites/201/2014/12/Nam_Saltzman_Krivokapic_Goldstein_PCC_2008.pdf.

Nespor, M., and I. Vogel. (1986) 2007. *Prosodic Phonology*. Dordrecht, the Netherlands: Foris.

Niemann, H., D. Mücke, H. Nam, L. Goldstein, and M. Grice. 2011. "Tones as Gestures: The Case of Italian and German." In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, 1486–1489. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/index.htm.

O'Dell, M. L., and T. Nieminen. 1999. "Coupled Oscillator Model of Speech Rhythm." In *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, edited by J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey, vol. 2, 1075–1078. New York: American Institute of Physics.

Oller, K. D. 1973. "The Effect of Position in Utterance on Speech Segment Duration in English." *Journal of the Acoustical Society of America* 54:1235–1247.

Pike, K. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Port, R. 1981. "Linguistics Timing Factors in Combination." *Journal of the Acoustical Society of America* 69:262–274.

Port, R. F., and T. van Gelder, eds. 1995. *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Pouplier, M. 2011. "The Atoms of Phonological Representations." In *The Blackwell Companion to Phonology*, edited M. van Oostendorp, K. Rice, B. Hume, and C. Ewen, 107–129. Hoboken, NJ: Wiley-Blackwell.

Prieto, P., and F. Torreira. 2007. "The Segmental Anchoring Hypothesis Revisited: Syllable Structure and Speech Rate Effects on Peak Timing in Spanish." *Journal of Phonetics* 35:473–500.

Prieto, P., J. van Santen, and J. Hirschberg. 1995. "Tonal Alignment Patterns in Spanish." *Journal of Phonetics* 23:429–451.

Proctor, M. 2009. "Gestural Characterization of a Phonological Class: The Liquids." PhD diss., Yale University.

Rakerd, B., W. Sennett, and C. A. Fowler 1987. "Domain-Final Lengthening and Foot Level Shortening in Spoken English." *Phonetica* 44:147–155.

Ramanarayanan, V., E. Bresch, D. Byrd, L. Goldstein, and S. Narayanan. 2009. "Analysis of Pausing Behavior in Spontaneous Speech Using Real-Time Magnetic Resonance Imaging of Articulation." *Journal of the Acoustical Society of America Express Letters* 126:EL160–EL165.

Ramanarayanan, V., D. Byrd, L. Goldstein, and S. Narayanan. 2010. "Investigating Articulatory Setting—Pauses, Ready Position, and Rest—Using Real-Time MRI." In *Proceedings of Interspeech 2010*, 1994–1997. https://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_1994.pdf.

Ramanarayanan, V., L. Goldstein, D. Byrd, and S. Narayanan. 2013. "A Real-Time MRI Investigation of Articulatory Setting across Different Speaking Styles." *Journal of the Acoustical Society of America* 134:510–519.

Rochet-Capellan, A., R. Laboissière, A. Galván, and J. L. Schwartz. 2008. “The Speech Focus Position Effect on Jaw-Finger Coordination in a Pointing Task.” *Journal of Speech Language and Hearing Research* 51:1507–1521.

Rusaw, E. 2013. “Modeling Temporal Coordination in Speech Production Using an Artificial Central Pattern Generator Neural Network.” PhD diss., University of Illinois.

Saltzman, E. L. 1986. “Task Dynamic Coordination of the Speech Articulators: A Preliminary Model.” In *Generation and Modulation of Action Patterns*, edited by H. Heuer and C. Fromm, 129–144. New York: Springer.

Saltzman, E. L. 1995. “Dynamics and Coordinate Systems in Skilled Sensorimotor Activity.” In *Mind as Motion: Explorations in the Dynamics of Cognition*, edited by R. F. Port and T. van Gelder, 149–173. Cambridge, MA: MIT Press.

Saltzman, E. L., and D. Byrd. 2000. “Task-Dynamics of Gestural Timing: Phase Windows and Multifrequency Rhythms.” *Human Movement Science* 19 (4): 499–526.

Saltzman, E. L., and J. A. S. Kelso. 1987. “Skilled Actions: A Task-Dynamic Approach.” *Psychological Review* 94:84–106.

Saltzman, E. L., and K. G. Munhall. 1989. “A Dynamical Approach to Gestural Patterning in Speech Production.” *Ecological Psychology* 1:333–382.

Saltzman, E., H. Nam, J. Krivokapić, and L. Goldstein. 2008. “A Task-Dynamic Toolkit for Modeling the Effects of Prosodic Structure on Articulation.” In *Proceedings of the Speech Prosody 2008 Conference*, edited by P. A. Barbosa, S. Madureira, and C. Reis, 175–184. http://www.isca-speech.org/archive/sp2008.

Selkirk, E. O. 1984. *Phonology and Syntax*: *The Relation between Sound and Structure*. Cambridge, MA: MIT Press.

Selkirk, E. O. 2011. “The Syntax-Phonology Interface.” In *The Handbook of Phonological Theory*, edited by J. Goldsmith, J. Riggle, and A. Yu, 435–484. Oxford: Blackwell.

Shattuck-Hufnagel, S., and A. Turk. 1996. “A Prosody Tutorial for Investigators of Auditory Sentence Processing.” *Journal of Psycholinguistic Research* 25:193–247.

Shattuck-Hufnagel, S., and A. Turk. 1998. “The Domain of Phrase-Final Lengthening in English.”, *Proceedings of the Sixteenth International Congress on Acoustics and 135th Meeting of the Acoustical Society of America*, Vol 2, 1235–1236.

Shattuck-Hufnagel, S. A., and A. Turk. 2011. “Durational Evidence for Word-Based vs. Prominence-Based Constituent structure in limerick speech.” In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, 1806–1809. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/index.htm.

Silverman, K., and J. Pierrehumbert. 1990. “The Timing of Prenuclear High Accents in English.” In *Papers in Laboratory Phonology I*, edited by J. Kingston and M. Beckman, 72–106. Cambridge: Cambridge University Press.

Snedeker, J., and E. Casserly. 2010. “Is It All Relative? Effects of Prosodic Boundaries on the Comprehension and Production of Attachment Ambiguities.” *Language and Cognitive Processes* 25:1234–1264.

Sproat, R., and O. Fujimura. 1993. “Allophonic Variation in English /1/ and Its Implications for Phonetic Implementation.” *Journal of Phonetics* 21:291–311.

Swerts, M. 1997. “Prosodic Features at Discourse Boundaries of Different Strength.” *Journal of the Acoustical Society of America* 101:514–521.

Tabain, M. 2003. “Effects of Prosodic Boundary on /aC/ Sequences: Articulatory Results.” *Journal of the Acoustical Society of America* 113:2834–2849.

Tilsen, S. 2009. “Multitimescale Dynamical Interactions between Speech Rhythm and Gesture.” *Cognitive Science* 33:839–879.

Turk, A. E., and J. R. Sawusch. 1997. “The Domain of Accentual Lengthening in American English.” *Journal of Phonetics* 25:25–41.

Turk, A., and S. Shattuck-Hufnagel, S. 2007. “Phrase-Final Lengthening in American English.” *Journal of Phonetics* 35:445–472.

Turk, A. E., and L. White. 1999. “Structural Influences on Accentual Lengthening.” *Journal of Phonetics* 27:171–206.

Turvey, M. T. 1990. “Coordination.” *American Psychologist* 45:938–953.

van Lieshout, P. 2004. “Dynamical Systems Theory and Its Application in Speech.” In *Speech Motor Control in Normal and Disordered Speech*, edited by B. Maassen, R. Kent, H. Peters, P. van Lieshout, and W. Hulstijn, 51–82. Oxford: Oxford University Press.

Wagner, M. 2005. “Prosody and Recursion.” PhD diss., MIT.

White, L. S. 2002. “English Speech Timing: A Domain and Locus Approach.” PhD diss., University of Edinburgh.

White, L., and A. E. Turk. 2010. “English Words on the Procrustean Bed: Polysyllabic Shortening Reconsidered.” *Journal of Phonetics* 38:459–471.

Wightman, C. W., S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price. 1992. “Segmental Durations in the Vicinity of Prosodic Phrase Boundaries.” *Journal of the Acoustical Society of America* 91:1707–1717.

Xu, Y. 1997. “Contextual Tonal Variations in Mandarin.” *Journal of Phonetics* 25:61–83.

Yi, H., and S. Tilsen. 2014. “A Gestural Account of Mandarin Tone Sandhi.” In *Proceedings of Meetings on Acoustics*, 060003. http://dx.doi.org/10.1121/2.0000029.

Commentary on Chapter 5: Time to Consider Nonoscillatory, Phonology-Extrinsic Timing Approaches to Prosody in Speech Motor Control

A. E. Turk

Introduction

Articulatory phonology/task dynamics (hereafter AP/TD) is the theory of speech production that provides the most comprehensive account of how prosodic structure influences the surface phonetic form of an utterance (Browman and Goldstein 1992; Browman and Goldstein 1995; Goldstein, Byrd, and Saltzman 2006; Saltzman and Munhall 1989; Nam, Goldstein, and Saltzman 2010; Saltzman et al. 2008). In this chapter, Krivokapić describes the mechanisms AP/TD uses to model coordination among speech articulators and a wide range of prosodic influences on coordination patterns. In this approach, prosodic gestures (*point-attractor oscillators*, i.e., damped oscillators that approximate a target and go no further) are used to model boundary-related lengthening, prominence-related lengthening, lexical tone, and pitch accent.

A second mechanism, a hierarchy of limit cycle planning oscillators (nondamped oscillators that actually oscillate), is used to model coordination and polysyllabic shortening. Because each gesture is assumed to be associated with a limit cycle planning oscillator, which freely oscillates, interoscillator coupling (entrainment) mechanisms can be used to model coordination among all types of prosodic gestures and segmental gestures and to model relationships among levels in a hierarchy of syllable, foot, and phrase oscillators. As Krivokapić shows, AP/TD is impressive in its coverage, and it is unrivaled among existing models of speech production in its accounts of temporal and prosodic phenomena.

The AP/TD view has been developed over many years and presented in many separate research articles, each dealing with a specific aspect of prosodic modeling. A major contribution of Krivokapić's essay is that it presents a synthesized view of the current AP/TD approach, drawn from a number of published sources. As a result, her review will be extremely useful to anyone working within the AP/TD approach, as well as to others working in other approaches who wish to understand it better.

In this commentary, I focus on one particular aspect of the theory: AP/TD's coupled oscillator approach to polysyllabic shortening (see next section), with some additional comments on the overall approach to speech timing. This aspect has been selected because it is a critical core assumption of the theory, which I argue is not compatible with a range of findings in the motor control literature. The additional comments address two further core assumptions: that gestures are coordinated with respect to their onsets, and that speech timing is controlled by phonology-intrinsic timing mechanisms. The first of these further assumptions is addressed in the "Challenges to AP/TD's Onset-Based Coordination" section, which presents evidence that motivates an alternative to onset-based coordination: coordination based on a goal-related part of movement (often the movement endpoint). The second is addressed in the “Challenges to AP/TD’s Phonology-Intrinsic Timing” section, which presents evidence that motivates an alternative to phonology-intrinsic timing: phonology-extrinsic, general-purpose timing mechanisms. The discussion in these two sections is necessarily short; see Turk and Shattuck-Hufnagel 2020) for a more detailed treatment.

Accounts of Poly-subconstituent Shortening: A Hierarchy of Syllable, Foot, and Phrase Oscillators versus a Means of Signaling Word Boundaries

As Krivokapić described, AP/TD’s approach to speech timing is based on gesture onset coordination via coupling between sets of planning oscillators for syllables, feet, and phrases. In this approach, the relative timing of gesture onsets is controlled via stable entrainment relationships (in-phase and antiphase) of the hierarchy of gestural coupling oscillators. The lengthening associated with prosodic prominence and boundaries is accounted for using π and μ_T gestures, which stretch the default gestural activation intervals with which they overlap.

The main motivation for the hierarchy of coupled syllable, foot, and phrase oscillators in AP/TD has come from findings of poly-segmental and polysyllabic shortening, where more subconstituents in a larger constituent (e.g., segments or syllables in a cross-word foot) results in durational compression of these segments or syllables. The duration of the larger constituent is shorter than expected based on the number of its subconstituent segments and/or syllables and their duration when produced in isolation. The constituent whose subconstituents undergo compression is often proposed to be a cross-word foot, that is, an interstress interval[1] that can, but does not have to, include multiple words or word fragments. On this view, for example, [*bake avo-*] in *bake avocadoes* and [*bake*] in *bake apples* would both be considered cross-word feet.

English and Swedish have been claimed to have polysyllabic shortening within cross-word feet. Campbell (1988, cited in Williams and Hiller 1994), Eriksson (1991), Williams and Hiller (1994), Kim and Cole (2005), and Kim (2006) have shown shorter stressed syllable durations when additional syllables follow the stressed syllable within a cross-word foot, even though the effect size is small (e.g., 10 to 15 percent per additional syllable, in Williams and Hiller’s 1994 study). Moreover, the total duration of the cross-word foot increases linearly with additional syllables and segments (see Dauer 1983 for a summary of such findings from the literature). Although compression does not yield cross-word foot isochrony, these findings suggest a *tendency* toward isochrony that might be due to periodic control at this level. As Classe (1939, 87) puts it, the surface shortening patterns might arise from a “rhythmic tendency,” which “has to contend with other factors which obscure its effects”; for example, as proposed in O’Dell and Nieminen (1999), Barbosa (2007), and AP/TD (Saltzman et al. 2008), from the interaction of coupled suprasegmental planning oscillators at multiple levels of prosodic constituency (syllable, cross-word foot, and phrase), which yields tendencies toward isochrony on the surface.

However, the interpretation of these findings as unambiguous evidence for syllable, cross-word foot, and phrase oscillators is not conclusive for two reasons. First, in most cases, the researchers did not control for the locations of word and/or phrase boundaries in their experimental materials. Thus, it is difficult to be sure whether apparent shortening in cross-word feet might be a spurious effect of word- or phrase-final lengthening. Word- and especially phrase-final syllables are known to be longer than non–word-final syllables. Thus, in corpus studies, it is possible that most monosyllabic cross-word feet correspond to single words that might exhibit word- or phrase-final lengthening; final lengthening might therefore account for the difference in duration between monosyllabic (e.g., *bake*) versus polysyllabic (e.g., *bake avo-*) cross-word feet, without requiring a separate oscillator-based polysyllabic shortening mechanism (Windmann, Šimko, and Wagner 2015).

Second, the interpretation of findings as evidence for syllable, foot, and phrase oscillators is made difficult because alternative possibilities for the constituents that might be responsible for the compression effects were often not considered. For example, whether polysyllabic shortening might operate in intervals other than cross-word feet, such as words or word-based constituents. With respect to this second question, Kim (2006) tested whether shortening occurs within words, rather than cross-word feet, for one of two speakers in the study, but did not test the possibility that it occurs within word-based content + function word units, as opposed to within interstress intervals. If the polysyllabic shortening effects described are spurious, then the main motivation for using suprasegmental oscillators to model such effects would be removed. Moreover, if polysyllabic shortening operates within intervals other than those proposed in AP/TD (that is, in units other than the syllable, cross-word foot, and phrase), then the theory would need to be modified to accommodate these different units.

Windmann, Šimko, and Wagner (2015) tested for these possibilities in a corpus study of 5.5 hours of automatically segmented broadcast speech produced by fifty-three speakers of British English. Like others in the literature, they found a clear compression effect of number of syllables on syllable duration in cross-word feet defined on the basis of word-level stress, especially for syllables that were phrasally prominent. However, this effect was due almost entirely to word-final lengthening; when the position of the measured syllable with respect to the word boundary was taken into account, that is, when separate analyses were conducted for word-final and nonfinal syllables, polysyllabic shortening effects within cross-word feet largely disappeared. This result suggests that compression mechanisms for syllables might not be needed as long as word- and phrase-final lengthening mechanisms are available.

A few effects consistent with polysyllabic shortening within cross-word feet nevertheless remained after position with respect to word boundaries was controlled for. For example, Windmann, Juraj, and Wagner (2015) found longer word-final syllables before a following word-initial stressed syllable than before a following word-initial unstressed syllable; this occurred both for unstressed syllables when preceded by a stressed syllable and for stressed syllables themselves. Effects of this type, which have been called *stress-adjacent lengthening effects*, have been found elsewhere, by, for example, Fowler (1977); also Rakerd, Sennett, and Fowler (1987); van Lancker, Kreiman, and Bolinger (1988); and Fant, Kruckenberg and Nord (1991), discussed in White (2014).

These effects are consistent with polysyllabic shortening within cross-word feet, but it is still unclear which type of constituent is responsible. If the unstressed syllable that followed the word boundary consisted of a function word (e.g., *a, us, the, him*), and the effect did not occur when the following unstressed syllable formed part of a following word, the results would be consistent with polysyllabic shortening within a word-based constituent that includes a following function word (sometimes called a clitic group) rather than within a cross-word foot that includes word fragments. It is true that, either way, these findings suggest that there are durational effects additional to boundary-related lengthening and prominence effects that need to be accounted for in some way. Critically, however, it is still unclear whether the observed effects involve cross-word feet that may contain word fragments, as proposed by AP/TD and Kim (2006), or whether they involve a word-based constituent such as the word-based clitic group.

One preliminary study that did test for the clitic group, word, and cross-word foot as possible domains of polysyllabic shortening found more support for word-based prosodic units (words and clitic groups) than for cross-word feet that may contain word fragments. Shattuck-Hufnagel and Turk (2011) contrasted these three types of candidate units in metrically regular poetic contexts (i.e., reciting limericks), where cross-word foot periodicity involving word fragments would be most likely to surface if they play a role in speech. They found that the rhyme interval durations of, for example, *bake* in *baking apples, bake us apples, bake an apple*, and *bake us an apple* were reliably shorter than *bake* in *bake apples*, for all three participants in their study. This result is consistent with polysyllabic shortening either within metrically defined inter-stress intervals (from the stressed onset syllable in *bake* to the stressed onset of *apples*), or with polysyllabic shortening within word-based clitic groups (including the content word *bake* plus the function words *us* and/or *an*).

The same study reported additional results that help to resolve the question of word-based versus metrically based constituents as the domain of polysyllabic shortening. Comparison of the rhyme interval duration of *bake* in, for example, *bake elixirs* and *bake avocadoes* versus in, for example, *baking, bake us, bake an*, and *bake us an* showed that the rhyme interval of *bake* was significantly shorter in the latter, word-based constituents. This suggests that word-based clitic groups were more influential than metrically based cross-word feet (such as *bake e-* and *bake avo-*). Only one of the three participants showed shorter rhyme interval durations for the metrically based constituents. These results suggest a stronger role for word-based constituents as compared to cross-word feet containing word fragments. They put polysyllabic shortening primarily in the domain of word-based structure, rather than of structure based on cross-word feet, suggesting that the timing control system based on syllables, feet, and phrases proposed in AP/TD would need to be modified to accommodate word-based constituents close in size to the lexical word.

A separate set of findings provides a further challenge to the use of oscillators to implement these shortening effects. White and Turk (2010) and Turk and Shattuck-Hufnagel (2000) found that polysyllabic shortening does not occur in all contexts; instead it occurs more often in phrasally stressed words, where it is also greatest in magnitude. These results challenge an oscillator-based implementation system because they suggest that even in the same utterances, not all words are affected by the suprasegmental compression mechanism. This reduces the motivation for a periodicity-based compression mechanism. Instead, these results support the idea that rather than being a reflection of a periodic control mechanism tending toward surface isochrony, polysyllabic shortening may be one of a set of mechanisms that speakers use to signal the locations of word-based prosodic constituent boundaries in an utterance (Turk 2012).

This section has described some of the challenges to AP/TD's oscillator-based approach to modeling systematic variation in speech timing that are raised by its account of poly-subconstituent shortening. Additional challenges to other aspects of oscillator-based timing control are described in the following two sections, which address, respectively, (i) the assumption of coordination based on movement onsets and (ii) the assumption of duration lengthening using phonology-intrinsic clock-slowing gestures.

Challenges to AP/TD's Onset-Based Coordination, by Evidence of Coordination Based on Movement Endpoints

AP/TD's interplanning-oscillator-coupling approach to coordination is based on gesture onsets. In this approach, the relative timing of gesture onsets is controlled via stable entrainment relationships (in-phase and antiphase) of gestural coupling oscillators. The lengthening associated with prominence and boundaries is accounted for using π and μ_T gestures, which stretch the default gestural activation intervals with which they overlap. This approach is challenged by findings suggesting that coordination can be based on parts of movement other than the onset, notably the part of movement that is most "behaviourally meaningful,"often the endpoint (Shaffer 1982; Semjen 1992). For example, Perkell and Matthies (1992) found that lip protrusion timing in relation to voicing onset was much more consistent at the end of the protrusion movement, as compared to a point near the beginning of protrusion. (See also Leonard and Cummins 2011 for related results for speech-accompanying gestures; Gentner, Grudin, and Conway 1980 for typing; Bootsma and van Wieringen 1990 and Katsumata and Russell 2012 for hitting balls; and Billon, Semjen, and Stelmach 1996 and Spencer and Zelaznik 2003 for finger-tapping.)

These findings do not find an obvious account in AP/TD because it does not have a representation of the time of gestural target approximation, that is, of mass-spring settling time. This is because in AP/TD the time of target approximation is an emergent property of mass-spring (gestural) stiffness. Although the time of gestural target approximation/mass-spring gestural settling corresponds to a fixed proportion of a gestural planning oscillator cycle at a default speaking rate, this correspondence changes when a μ_T gesture stretches the gestural planning oscillator cycles at particular locations in an utterance, and at fast speaking rates, when gestural planning oscillator cycles are shortened. Findings of movement-endpoint–based coordination motivate alternative approaches that make it possible to refer to the (surface) timing of movement endpoints, and in which endpoint times serve as reference points for coordination (Shaffer 1982, Semjen 1992). For example, Lee's (1998, 2009) general tau theory proposes that endpoint coordination can be accomplished if movement tau, that is, the time remaining to movement endpoint attainment at the current movement rate, is kept in constant proportion either to the tau of another movement, or to the tau of an internally generated tau guide. This mechanism ensures synchronous endpoint achievement, as long as two movements (or a movement with the tau guide) are tau-coupled before the end of the movement.

In sum, evidence for movement coordination at goal-related parts of movements, e.g., endpoints, presents challenges to the onset-based coordination mechanisms of AP/TD. The second set of challenges is raised by AP/TD's use of a phonology-intrinsic clock rather than a general-purpose timekeeping mechanism.

Challenges to AP/TD's Phonology-Intrinsic Timing, by Evidence for General-Purpose Timekeeping Mechanisms

This section discusses the fact that AP/TD's account of prosodically governed duration lengthening relies on the slowing of a phonology-intrinsic clock by π/μ_T gestures and contrasts this mechanism with the use of general-purpose, phonology-extrinsic timekeeping mechanisms that meter out units of solar (i.e., surface) time. As noted herein and as described by Krivokapić, in AP/TD the lengthening associated with prominence and boundaries is accounted for using π and μ_T gestures, which stretch the default gestural activation intervals with which they overlap in time. This reliance on the adjustment of default timing is challenged by several sets of findings, including (i) findings showing that timing variability is greater for longer duration intervals, (ii) findings relating to prominence and boundaries suggestive of constraints on surface durations, and

(iii) the fact that π and μ_T gestures warp phonological time in relation to solar time, in a nonlinear manner across an utterance.

In relation to the first of these findings, many investigators have shown that timing variability for a movement grows with its duration. Results of this kind have been reported for nonspeech movements by Treisman (1963); Gibbon (1977); Schmidt et al. (1979); Rosenbaum and Patshnik (1980a, 1980b); Wing (1980); Hancock and Newell (1985); Wearden (1991); Ivry and Corcos (1993); Ivry and Hazeltine (1995); Spencer and Zelaznik (2003); Merchant et al. (2008); and Merchant, Zarco, and Prado (2008); see Malapani and Fairhurst (2002) for a review. For speech movements, see studies by Byrd and Saltzman (1998); Edwards, Beckman and Fletcher (1991); Remijsen and Gilley (2008); Chen (2006); Nakai et al. (2012); and Lefkowitz (2017).

These findings are challenging for AP/TD, because in that framework, activation intervals are stretched not by adding extra time to the activation intervals, but instead by a "clock slowing" mechanism. That is, the μ_T gestures slow the planning + suprasegmental ensemble clock during intervals overlapped by the μ_T gesture. In this approach, an interval that has been stretched by a μ_T gesture therefore has the same number of AP/TD planning + suprasegmental ensemble clock units as a corresponding interval that has not been stretched by a π or μ_T gesture. The π/μ_T approach therefore does not provide an account of the greater variability for longer surface duration intervals, because these longer surface duration intervals are not longer in AP/TD planning + suprasegmental ensemble clock time. The findings appear to instead require the representation of surface durations in solar time units, where longer duration intervals contain more surface time intervals (that is, they are longer in solar time) than shorter duration intervals. Greater variability of these longer duration intervals can be explained by a "noisy" timekeeper, in which variability correlates with interval duration (Gallistel 1999; Gallistel and Gibbon 2000; Jones and Wearden 2004; Shouval et al. 2014). It is not clear how AP/TD can account for these findings, because surface durations are only emergent in this approach, and are not represented.

With respect to the second point, findings suggestive of constraints on surface durations in prominent and boundary-adjacent positions provide another challenge to the use of π/μ_T gestures. For example, in Northern Finnish, a quantity language in which short versus long quantities are not signaled by vowel quality but by vowel duration, phonemically short vowels show smaller amounts of prominence- and boundary-related lengthening than phonemically long vowels do (Nakai et al. 2009; Nakai et al. 2012; Remijsen and Gilley 2008). These findings suggest that speakers of these languages avoid large amounts of lengthening on phonemically short vowels to maintain their surface durational distinction with phonemically long vowels. These findings are difficult to explain if (as in AP/TD) surface durations cannot be represented, because prominence- and boundary-related lengthening are emergent properties of default activation interval durations plus π/μ_T adjustments, without the possibility of representing the surface duration results as a goal.

With respect to the third point, AP/TD's slowing of the phonology-intrinsic clock in prosodically prominent and phrase-final positions, and at slow speaking rates, results in a lack of linear correspondence between AP/TD's planning + suprasegmental clock time and solar time. Because AP/TD does not make use of phonology-extrinsic timekeepers, this "time-warping" may have unintended and undesirable consequences. This may occur, for example, in cases where singers or speakers need to interact with external events, as when singing to instrumental music (where matching events in solar time is critical) or timing an oral presentation to end at a particular point in solar time.

Taken together, these findings present a challenge to AP/TD's use of phonology-intrinsic timing mechanisms and motivate the development of models in which speech timing is controlled instead by general-purpose timing mechanisms that are extrinsic to the phonology.

Conclusion

While AP/TD has successfully modeled many aspects of segmental coordination and prosody, as demonstrated in Krivokapić's essay, there are a number of lines of evidence that are difficult to reconcile with its coupled oscillator approach to coordination; (i) its default-adjustment approach to longer durations at prosodic boundaries and in prominent positions; and (ii) its system of syllable-, cross-word foot- and phrase-level oscillators. These observations highlight the importance of developing alternative models of speech production and testing the predictions of these models against predictions of AP/TD. Turk and Shattuck-Hufnagel (2020) provide additional evidence that supports phonology-extrinsic timing-based approaches. Considerable work will be required to bring such models to the state of experimentally tested implementation that has made the AP/TD approach so influential and widely discussed.

Note

1. Note that the term *cross-word foot* does not specify the type of stress foot, that is, the level of prominence that delimits the foot. For example, Abercrombie's (1991) cross-word feet were delimited by "chest pulses" and were envisioned to be delimited by phrasal prominences. In AP/TD, cross-word feet are defined on the basis of word-level stress. See also Williams and Hiller (1984), who performed tests of polysyllabic shortening within feet of different types.

References

Abercrombie, D. 1991. "Some Functions of Silent Stress." In *Fifty Years in Phonetics*, edited by D. Abercrombie. Edinburgh: Edinburgh University Press.

Barbosa, P. A. 2007. "From Syntax to Acoustic Duration: A Dynamical Model of Speech Rhythm Production." *Speech Communication* 49 (9): 725–742.

Billon, M., A. Semjen, and G. E. Stelmach. 1996. "The Timing Effects of Accent Production in Periodic Finger-Tapping Sequences." *Journal of Motor Behavior* 28 (3): 198–210.

Bootsma, R. J., and P. C. van Wieringen. 1990. "Timing an Attacking Forehand Drive in Table Tennis." *Journal of Experimental Psychology: Human Perception and Performance* 16 (1): 21–29.

Browman, C. P., and L. Goldstein. 1992. "Articulatory Phonology: An Overview." *Phonetica* 49 (3–4): 155–180.

Browman, C. P., and L. Goldstein. 1995. "Dynamics and Articulatory Phonology." In *Mind as Motion*, edited by R. F. Port and T. v. Gelder, 175–193. Cambridge, MA: MIT Press.

Byrd, D., and E. Saltzman. 1998. "Intragestural Dynamics of Multiple Prosodic Boundaries." *Journal of Phonetics* 26:173–199.

Campbell, W. N. 1988. "Foot-Level Shortening in the Spoken English Corpus." Paper presented at the Seventh Forensic Anthropology Society of Europe Symposium, Edinburgh, UK, September 1988.

Chen, Y. 2006. “Durational Adjustment under Corrective Focus in Standard Chinese.” *Journal of Phonetics* 34 (2): 176–201.

Classe, A. 1939. *The Rhythm of English Prose*. Oxford: Basil Blackwell.

Dauer, R. M. 1983. “Stress-Timing and Syllable-Timing Re-Analysed.” *Journal of Phonetics* 11:51–62.

Edwards, J., M. E. Beckman, and J. Fletcher. 1991. “The Articulatory Kinematics of Final Lengthening.” *Journal of the Acoustical Society of America* 89 (1): 369–382.

Eriksson, A. 1991. “Aspects of Swedish Speech Rhythm.” PhD diss., University of Gothenburg.

Fant, G., A. Kruckenberg, and L. Nord. 1991. “Stress Patterns and Rhythm in the Reading of Prose and Poetry with Analogies to Music Performance.” Paper presented at the Music, Language, Speech, and Brain International WennerGren Symposium, Stockholm.

Fowler, C. A. 1977. *Timing Control in Speech Production*. Bloomington: Indiana University Linguistics Club.

Gallistel, C. R. 1999. “Can a Decay Process Explain the Timing of Conditioned Responses?” *Journal of the Experimental Analysis of Behavior* 71:264–271.

Gallistel, C. R., and J. Gibbon. 2000. “Time, Rate, and Conditioning.” *Psychological Review* 107 (2): 289–344.

Gentner, D. R., J. Grudin, and E. Conway. 1980. *Finger Movements in Transcription Typing*. Technical Report 8001. San Diego: Center for Human Information Processing, University of California, San Diego.

Gibbon, J. 1977. “Scalar Expectancy Theory and Weber’s Law in Animal Timing.” *Psychological Review* 84 (3): 279–325.

Goldstein, L., D. Byrd, and E. Saltzman. 2006. “The Role of Vocal Tract Gestural Action Units in Understanding the Evolution of Phonology.” In *Action to Language via the Mirror Neuron System*, edited by M. A. Arbib, 215–249. Cambridge: Cambridge University Press.

Hancock, P. A., and K. M. Newell. 1985. “The Movement Speed-Accuracy Relationship in Space-Time.” In *Motor Behavior: Programming, Control, and Acquisition*, edited by Herbert Heuer, U. Kleinbeck, and K.-H. Schmidt, 153–185. Berlin: Springer-Verlag.

Ivry, R., and D. M. Corcos. 1993. “Slicing the Variability Pie: Component Analysis of Coordination and Motor Dysfunction.” In *Variability and Motor Control*, edited by K. Newell and D. Corcos, 415–447. Champaign, IL: Human Kinetics Publishers.

Ivry, R. B., and R. E. Hazeltine. 1995. “Perception and Production of Temporal Intervals across a Range of Durations: Evidence for a Common Timing Mechanism.” *Journal of Experimental Psychology—Human Perception and Performance* 21 (1): 3–18.

Jones, L. A., and J. H. Wearden. 2004. “Double Standards: Memory Loading in Temporal Reference Memory.” *Quarterly Journal of Experimental Psychology* 57B (1): 55–77.

Katsumata, H., and D. M. Russell. 2012. “Prospective versus Predictive Control in Timing of Hitting a Falling Ball.” *Experimental Brain Research* 216 (4): 499–514.

Kim, H. 2006. “Speech Rhythm in American English: A Corpus Study.” PhD diss., University of Illinois at Urbana-Champaign.

Kim, H., and J. Cole. 2005. “The Stress Foot as a Unit of Planned Timing: Evidence from Shortening in the Prosodic Phrase.” In *Proceedings of Interspeech/Eurospeech 2005*, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4–8, 2005; ISCA Archive. http://www.isca-speech.org/archive/interspeech_2005.

Lee, D. N. 1998. “Guiding Movement by Coupling Taus.” *Ecological Psychology* 10 (3–4): 221–250.

Lee, D. N. 2009. “General Tau Theory: Evolution to Date.” *Perception* 38:837–858.

Lefkowitz, L. M. 2017. “Maxent Harmonic Grammars and Phonetic Duration.” PhD diss., UCLA.

Leonard, T., and F. Cummins. 2011. “The Temporal Relation between Beat Gestures and Speech.” *Language and Cognitive Processes* 26 (10): 1457–1471.Malapani, C., and S. Fairhurst. 2002. “Scalar Timing in Animals and Humans.” *Learning and Motivation* 33:156–176.

Merchant, H., W. Zarco, R. Bartolo, and L. Prado. 2008. “The Context of Temporal Processing Is Represented in the Multidimensional Relationships between Timing Tasks.” *PLoS One* 3 (9): e3169.

Merchant, H., W. Zarco, and L. Prado. 2008. “Do We Have a Common Mechanism for Measuring Time in the Hundreds of Millisecond Range? Evidence from Multiple-Interval Timing Tasks.” *Journal of Neurophysiology* 99:939–949.

Nakai, S., S. Kunnari, A. Turk, K. Suomi, K., and R. Ylitalo. 2009. “Utterance-Final Lengthening and Quantity in Northern Finnish.” *Journal of Phonetics* 37:29–45.

Nakai, S., A. E. Turk, K. Suomi, S. C. Granlund, R. Ylitalo, and S. Kunnari,. 2012. “Quantity Constraints on the Temporal Implementation of Phrasal Prosody in Northern Finnish.” *Journal of Phonetics* 40:796–807.

Nam, H., L. Goldstein, and E. Saltzman. 2010. “Self-Organization of Syllable Structure: A Coupled Oscillator Model.” In Approaches to Phonological Complexity, edited by François Pellegrino, E. Marisco, and I. Chitoran, 99–328. Berlin: Mouton de Gruyter.

O’Dell, M. L., and T. Nieminen. 1999. “Coupled Oscillator Model of Speech Rhythm.” In *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, edited by J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey, vol. 2, 1075–1078. College Park, MD: American Institute of Physics.

Perkell, J. S., and M. L. Matthies. 1992. “Temporal Measures of Anticipatory Labial Coarticulation for the Vowel /u/: Within-Subject and Cross-Subject Variability.” *Journal of the Acoustical Society of America* 91 (5): 2911–2925.

Rakerd, B., W. Sennett, and C. A. Fowler. 1987. “Domain-Final Lengthening and Foot-Level Shortening in Spoken English.” *Phonetica* 44:47–155.

Remijsen, B., and L. Gilley. 2008. “Why Are Three-Level Vowel Length Systems Rare? Insights from Dinka (Luanyjang Dialect).” *Journal of Phonetics* 36:318–344.

Rosenbaum, D. A., and O. Patashnik. 1980a. “A Mental Clock Setting Process Revealed by Reaction Times.” In *Tutorials in Motor Behavior*, edited by G. E. Stelmach and J. Requin, 487–499. Amsterdam: North-Holland.

Rosenbaum, D. A., and O. Patashnik. 1980b. “Time to Time in the Human Motor System.” In *Attention and Performance*, vol. 8, edited by R. S. Nickerson. Hillsdale, NJ: Erlbaum.

Saltzman, E. L., and K. G. Munhall. 1989. “A Dynamical Approach to Gestural Patterning in Speech Production.” *Ecological Psychology* 1 (4): 333–382.

Saltzman, E., H. Nam, J. Krivokapić, and L. Goldstein. 2008. “A Task-Dynamic Toolkit for Modeling the Effects of Prosodic Structure on Articulation.” In *Proceedings of the Speech Prosody 2008 Conference*, edited by P. A. Barbosa, S. Madureira, and C. Reis, 175–184.

Schmidt, R. A., H. Zelaznik, B. Hawkins, J. S. Frank, and J. T. Quinn. 1979. “Motor-Output Variability: A Theory for the Accuracy of Rapid Motor Acts.” *Psychological Review* 86 (5): 415–451.

Semjen, A. 1992. “Determinants of Timing in Serial Movements.” In *Time, Action and Cognition*, vol. 66, edited by F. Macar, V. Pouthas, and W. J. Friedman, 247–261. Berlin: Springer.

Shaffer, L. H. 1982. “Rhythm and Timing in Skill.” *Psychological Review* 89 (2): 109–122.

Shattuck-Hufnagel, S., and A. Turk. 2011. “Durational Evidence for Word-Based vs. Prominence-Based Constituent Structure in Limerick Speech.” In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*. Hong Kong. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/index.htm.

Shouval, H. Z., M. G. Hussain Shuler, A. Agarwal, and J. P. Gavornik. 2014. “What Does Scalar Timing Tell Us about Neural Dynamics?” *Frontiers in Human Neuroscience* 8:438.

Spencer, R. M. C., and H. N. Zelaznik. 2003. “Weber (Slope) Analyses of Timing Variability in Tapping and Drawing Tasks.” *Journal of Motor Behavior* 35 (4): 371–381.

Treisman, M. 1963. “Temporal Discrimination and the Indifference Interval: Implications for a Model of the Internal Clock.” *Psychological Monographs* 77 (13): 1–31.

Turk, A. 2012. “The Temporal Implementation of Prosodic Structure.” In *The Oxford Handbook of Laboratory Phonology*, edited by A. C. Cohn, C. Fougeron, and M. K. Huffman, 242–253. New York: Oxford University Press.

Turk, A. E., and S. Shattuck-Hufnagel. 2000. “Word-Boundary-Related Duration Patterns in English.” *Journal of Phonetics* 28:397–440.

Turk, A., and S. Shattuck-Hufnagel. 2020. *Speech Timing: Implications for Theories of Phonology, Phonetics, and Speech Motor Control*. New York: Oxford University Press.

Van Lancker, D., J. Kreiman, and D. Bolinger. 1988. “Anticipatory Lengthening.” *Journal of Phonetics* 61:339–347.

Wearden, J. H. 1991. “Do Humans Possess an Internal Clock with Scalar Timing Properties?” *Learning and Motivation* 22:59–93.

White, L. 2014. “Communicative Function and Prosodic Form in Speech Timing.” *Speech Communication* 63–64:38–54.

White, L., and A. E. Turk. 2010. “English Words on the Procrustean Bed: Polysyllabic Shortening Reconsidered.” *Journal of Phonetics* 38:459–471.

Williams, B., and S. M. Hiller. 1994. “The Question of Randomness in English Foot Timing: A Control Experiment.” *Journal of Phonetics* 22:423–439.

Windmann, A., J. Šimko, and P. Wagner. 2015. “Polysyllabic Shortening and Word-Final Lengthening in English.” In *Proceedings of Interspeech* 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6–10; ISCA Archive. http://www.isca-speech.org/archive/interspeech_2015.

Wing, A. M. 1980. “The Long and Short of Timing in Response Sequences.” In *Tutorials in Motor Behavior*, edited by G. E. Stelmach and J. Requin, 469–484. Amsterdam: North-Holland.

6 The Trouble with ToBI

D. R. Ladd

6.1 Introduction

ToBI came into existence about 1990. Its original stated purpose was to serve as an agreed standard for the prosodic annotation of English speech databases. The primary aim of such annotation was to render databases easily searchable—to make it possible to look for all the phrase-final rises, or all the low accented syllables, or all the sequences of downstepped accents. Two main groups of researchers were behind the initiative: a group associated with Janet Pierrehumbert's tone-based autosegmental analysis of English intonation contours (e.g., Pierrehumbert 1980; Beckman and Pierrehumbert 1986; Pierrehumbert and Hirschberg 1990) and a group who had been transcribing hierarchical prominence and phrasing structure using numerical notations of boundary strength ("break indices") at each orthographic word boundary (Price et al. 1991; Wightman et al. 1992). Together the autosegmental tones and the numerical break indices yielded the acronym ToBI.

The early consultation meetings involved not only representatives of the two groups but a number of intonation researchers associated with the British and Dutch traditions, as well as several speech technologists not associated with any particular school of prosodic analysis. At these meetings, there was general agreement on the value of a common transcription standard, but little consensus on how to achieve it. There were long discussions of specific proposals, some details of which I summarize in the description of ToBI presented in Ladd (2008a, 104–113), and one important aspect of which I consider at greater length in the next section. Eventually a published standard resulted (Silverman et al. 1992). There were some limited evaluation studies (Pitrelli, Beckman, and Hirschberg 1994), and extensive training materials were created (Beckman and Ayers Elam 1997; Brugos, Shattuck-Hufnagel, and Veilleux 2006). The various components of a full ToBI annotation are described in more detail in the historical and theoretical summary by Beckman, Hirschberg, and Shattuck-Hufnagel (2005).

As early as the mid-1990s, however, in a kind of "mission creep," the ToBI tonal analysis took on a life of its own separate from the break-index analysis, and it rapidly became common to regard a ToBI tone sequence as the most acceptable way of reporting intonational data in published articles that had nothing to do with the labeling of speech databases (e.g., Birch and Clifton 1995). My comments in this chapter focus on this general use of the tonal notation and its claims to broader validity. To avoid any confusion about my intentions, I emphasize that my critique deals specifically with ToBI in its institutionalized form and with the theoretical justifications for that form laid out by Beckman, Hirschberg, and Shattuck-Hufnagel (2005). I remain strongly convinced of the potential of the autosegmental-metrical (AM) approach to describing intonation, concisely yet comprehensively presented in Arvaniti (chapter 1, this volume). The trouble with ToBI, in my view, is that it precludes exploration of some of the important theoretical issues within the general AM approach.

6.2 Phonological Distinctions and Phonetic Categories

As the original ToBI tonal transcription system acquired the status of received analysis, there was a proliferation of ToBI notation systems for other languages (e.g., Japanese, Venditti 1997; German, Grice et al. 1996; Greek, Arvaniti and Baltazani 2005); more recently, there has been a move to create a universally applicable international prosodic alphabet (Hualde and Prieto 2016). This dual outcome—specific ToBI variants for specific languages, followed by a call for a transcription system that is not specific to any language—reflects the tension that surfaced during the first consultation meetings between what we might call “phonological” and “phonetic” conceptions of transcription. This is a fundamental tension in any symbolic annotation system: What is the proper basis for choosing one symbolic phonetic string over another? Even in segmental transcription with the International Phonetic Alphabet (IPA), where the issues are more clearly understood, there is no agreement; the problem is usually treated as a matter of “broad” and “narrow” transcription. In a narrow transcription, one is theoretically annotating all the phonetic detail one can perceive, but everyone acknowledges that there are degrees of narrowness. In a broad transcription, IPA symbols are used (e.g., in dictionaries) to spell the phonemically distinct segment types in any given language. Because of the phonemic principle, such broad transcriptions of individual words are often uncontroversial, but even so, the line between broad and narrow remains fuzzy.

For one thing, there is a practical issue: different varieties of the same language often have slightly different systems of phonemic contrast. In a word like *Marrakesh*, the initial vowel is systemically different for speakers who pronounce *Mary* and *marry* identically and speakers who distinguish them. Such cases are fairly well understood, and it is possible (albeit ultimately arbitrary) to establish transcription policies to deal with them. In theory, the ideal solution might be to give different transcriptions for different varieties on the basis of the phonemic distinctions they actually make, but compromises are usually adopted to avoid an impractical degree of fragmentation into multiple transcription systems. This issue is extensively discussed in the Wikipedia guidance for the phonetic transcription of English (https://en.wikipedia.org/wiki/Help:IPA/English), which takes its overall aim to be that of maintaining a unitary transcription system for all of English wherever possible. I return to this problem (and Wikipedia’s solution) in the next section.

But there is another more fundamental problem for which ordinary notions of phonemic distinctiveness provide no solution. This is the problem of salient subphonemic detail, especially in connected speech. In English, these phenomena include such things as cluster simplification and place assimilation (Do we transcribe the sentence *It costs ten pounds* as /ɪtˈkɒstsˈtɛnˈpaʊndz/ or /ɪʔˈkɒsːˈtɛmˈpaʊnz/?), but the problem can be simply illustrated even in isolated words by the use of the glottal stop. In many North American varieties of English, orthographic T is regularly realized by a glottal stop before sonorants in names like *Atwood* and *Whitlock*, and before unstressed /ən/ or syllabic /n̩/ in names like *Clinton* and *Horton* and words like *blatant* and *important*. Should this difference be noted in transcribing? Superficially, these cases seem to be covered by the phonemic principle: the distribution of glottal stop is phonologically conditioned; glottal stop is therefore an allophone of /t/ and should not be indicated in a phonemic transcription. Somehow, though, some allophones are more noticeable than others.

When Bill Clinton was elected U.S. president in 1992, for example, I heard a British television reporter comment with some surprise on the fact that most Americans pronounced his name with a glottal stop. Given the well-known stigma attached to many uses of glottal stop in British English, it is obviously not surprising that a British reporter noticed its nonstigmatized use in American pronunciations of *Clinton*. But my point here is not sociolinguistic; it is theoretical. What kind of entity is glottal stop in English? The expectation of classical phonemic theory is that phonetic distinctions that can be used to distinguish different meanings in a given situation will be noticed and reliably reproduced by the language's native speakers, while phonetic distinctions that are completely predictable—allophonic differences—will be reliably reproduced but will not be noticed. This expectation is met in some cases, as, for example, the different aspiration of English voiceless stops in absolute syllable initial position and following /s/. This difference predictably astonishes beginning linguistics students when it is used to introduce the phonemic principle. But in other cases—English glottal stop, the German [ç/χ] distinction (Moulton 1947), the Russian [i/ɯ] (и/ы) distinction (Kiparsky 2014)—what we might call "allophonic awareness" is undeniably real. Somehow the members of these pairs constitute distinct phonetic categories for many speakers of these languages. If these categories cannot be grounded in phonemic contrast, where do they come from?

I do not propose to try to answer this question here, but simply to suggest that it lies at the heart of the difficulties with ToBI (and ultimately with the IPA; see, e.g., Pike 1943, 42–55; Ashby et al. 1995). The notion of phonological distinctiveness, and the idea that different languages make different distinctions, is well established and reasonably well understood. The goal of transcribing only the phonological distinctions in the intonation system of a given language variety, carefully laid out for ToBI by Beckman, Hirschberg, and Shattuck-Hufnagel (2005), is theoretically coherent. But there is also a kind of strictly phonetic distinctiveness that is poorly theorized yet intuitively obvious—the kind of distinctiveness that tells us that /t/ and glottal stop are different. This is what leads almost anyone who sets about transcribing intonation to question the theoretical discipline promoted by Beckman, Hirschberg, and Shattuck-Hufnagel and to try to find symbolic means to represent audible differences that have no obvious place in the official phonology.

In the following sections, I discuss reasons why I believe Beckman, Hirschberg, and Shattuck-Hufnagel's confidence in ToBI's phonological foundations is at best premature.

6.3 ToBI and the Phonological Analysis of Intonation

The first question that ToBI or any other transcription standard needs to address is what to do about differences between different language varieties. The trend in ToBI, sketched by Beckman, Hirschberg, and Shattuck-Hufnagel (2005), has been to narrow the range of varieties of English that the system claims to cover. The original idea was that ToBI was to be a suitable transcription for intonation in at least Southern British, Australian, and North American English. This list implicitly acknowledges that Irish and Northern British intonation is conspicuously different in some ways, but treats the rest of "English" as involving more or less the same system of intonational phonology. But by the late 1990s, the ToBI system being elaborated by various scholars at Ohio State, MIT, and elsewhere in the United States was renamed MAE-ToBI and was said to cover Mainstream American English.

I see at least two reasons why this narrowing is a bad idea. The first is merely practical: given the high degree of uniformity and mutual intelligibility across most varieties of "English," there are obvious advantages to having a unified transcription. For segmental transcription, this goal is explicitly adopted by the Wikipedia guidelines for the use of the IPA to indicate pronunciation. This obviously requires some compromises, mostly in the direction of indicating phonological distinctions made by large and/or sociolinguistically important groups of English speakers even where those same distinctions are neutralized by other equally significant groups. For example, Wikipedia recommends transcribing the presence or absence of postvocalic /r/ in English words such as *doorway* and *daughter*, even though the initial syllables are pronounced identically by millions of English speakers. (In terms of the standard Wells lexical sets, that is, the recommendation is to distinguish between THOUGHT and FORCE/NORTH). This recommendation applies even to place names in nonrhotic parts of the English-speaking world, such as *Worcester* (transcribed /ˈwʊstər/) and *Melbourne* (transcribed /ˈmɛlbərn/), which would normally be pronounced locally without the final /r/. In the same way, to repeat an example used earlier, Wikipedia transcribes the Anglicized pronunciation of *Marrakesh* as /ˌmærəˈkɛʃ/, observing the distinction between the vowels of TRAP and SQUARE in transcribing the first syllable even though the majority of North American speakers of English do not maintain this distinction before /r/. While such compromises are arbitrary—for example, Wikipedia explicitly recommends ignoring the distinction between NORTH and FORCE made by a relatively small minority of English speakers in Scotland, Ireland, and parts of the United States—they offer the convenience of a uniform transcription system that is fairly easy to apply consistently and interpretably to the pronunciation of hundreds of millions of speakers. In turning away from a similar goal, the developers of MAE-ToBI ostensibly chose strict theoretical adherence to the ideal of transcription based on an analysis of the system of phonological contrasts.

Yet this adherence is illusory, not least because "Mainstream American English" is ultimately the same kind of abstraction as "English" is, namely, an abstraction over a set of idiolects. And in any case, there is a more fundamental problem with claiming that MAE-ToBI is properly based on the phonological contrasts of only one variety of English. This claim presupposes that we have figured out how intonational phonology works well enough to establish what those phonological contrasts are. In my view, our understanding of intonational phonology is actually still fairly primitive, and a standard transcription that purports to be based on a correct phonological analysis prematurely closes off avenues of investigation and theoretical debate. A clear illustration of this problem involves the analysis of phrase-final level pitch. This is seen in a variety of contexts in English, perhaps the most familiar being the stereotypical contour used for calling someone from a distance (e.g., Gibbon 1976; Leben 1976; Ladd 1978; Gussenhoven 1993). In this contour, the stressed syllable is produced on a high level pitch, and the remaining syllables are produced on another level that is often three semitones lower:

Ma—
ry—

In Pierrehumbert's (1980) original analysis of this phenomenon, the level pitch on the final syllable was analyzed phonologically as a combination of a high phrase accent (H-) followed by a low boundary tone (L%). Though this transcription might appear to indicate a contour that falls in pitch from a postnuclear high, the analysis included the proviso that the L% is "upstepped" in this specific context and is realized at the level of the phrase accent. This contrasts with the phrase-final sequence H- H%, where the upstepping of the H% leads to a rise. This analysis works, in the sense that it can be made to generate the contrast between rising and sustained level phrase-final pitch, and it remains the official MAE-ToBI transcription of contours involving sustained final level.

However, the idea of the upstepped L% was subject to criticism during the burst of theoretical discussion and development in the 1980s and 1990s that followed in the wake of Pierrehumbert (1980). One line of theoretical criticism dealt with the status of phrase accents and boundary tones in general. In Ladd (1983), I proposed that boundary tones need not be part of a well-formed tonal string and argued that the absence of a final rise or fall in phrases ending on steady level pitch was best seen as a reflection of the absence of a boundary tone. This idea was subsequently adopted by Gussenhoven (1984) and Grabe (1998) on the basis of evidence from English, Dutch, and German; it was later incorporated into the ToDI system for transcribing Dutch intonation (Gussenhoven et al. 2003), though not into the GToBI system for German (Grice et al. 1996; Grice and Baumann 2002). Another line of theoretical criticism questioned the appropriateness of opaque phonological analyses that depend on the use of phonetic realization rules like upstep and downstep to yield surface pitch contours. For example, in Ladd (1983), I argued against Pierrehumbert's (1980) original phonological analysis of downstep, which relied on abstract tonal strings subject to rather complex phonetic realization rules (see Arvaniti, chapter 1, section 1.2.4, this volume), and I suggested instead that downstep is an independent phonological choice that can apply to any high accent. As it happened, Pierrehumbert's abstract analysis of downstep was an early casualty of the consultations that led to ToBI, being replaced in the agreed standard by a transparent "downstep feature" transcribed with the traditional Africanist device /!/, but the comparably abstract H- L% analysis of level phrase-final pitch was retained unchanged.

The most specific critique of the H- L% analysis of level phrase-final pitch (Ladd 1983; Mayo, Aylett, and Ladd 1997) focused on the fact that it leaves no obvious way of describing numerous British English intonation patterns in which the pitch gradually falls, or "slumps" after a nuclear accentual rise (the "rise-plateau-slump" contour of Cruttenden 1994). The two issues summarized in the preceding paragraph—the status of boundary tones and the abstractness of tonal strings—are general theoretical questions that it might have been useful to debate more thoroughly while the potential of the general AM approach to describing intonation was being explored. But these questions went largely unexamined, because it was possible for the Pierrehumbert/ToBI analysis to focus on the argument from the rise-plateau-slump and to shelter behind the claim that ToBI deals only with Mainstream American English. ToBI's theoreticians came to favor the view that different analyses reflect genuine differences between language varieties:

> A ToBI framework system devised for one language variety cannot be assumed to be applicable even to other varieties of the same language, without first establishing appropriate intonational and prosodic analyses for each variety. Any claim that the symbolic tags [i.e., ToBI labels] are comparable across varieties must be based on a thorough variety-specific analysis of each of the varieties involved. (Beckman, Hirschberg, and Shattuck-Hufnagel 2005, 42)

This is clearly an intellectually respectable position. The phonological distinctions of what is sometimes termed "Urban North British" varieties of English, to say nothing of Dutch and German, are obviously irrelevant to determining the phonological contrasts of Mainstream American. But they are not so easily dismissed as irrelevant to a broader theoretical understanding of the nature of boundary tones or the scope and power of the phonetic realization rules involved in intonational phonology. I believe that the determination to hold firm to an agreed standard based on one specific autosegmental analysis of one specific variety of English has impeded the full discussion of important theoretical issues that arise from the general AM approach.

6.4 ToBI and Gradience

The question of the upstepped boundary tone is a mere detail, however, compared to what I see as the central unsolved problem with intonational phonology. This is the phenomenon that Bolinger (1961) termed "gradience." In the ordinary lexicon, word meanings are pretty much discrete, and if two words are different, we have reasonably well-established ways of describing the phonetic properties that signal the difference. The subtlety of the phonetic distinctions doesn't matter: *ankle* and *uncle* are similar enough phonetically to cause problems for many second-language speakers of English, but we know they mean something different and can confidently conclude that the phonetic difference between their initial vowels is a systematic part of English segmental phonology.

We can't proceed so confidently with intonation, because our metalanguage for describing intonational phonetics is not so well developed. And that, I believe, is due to the fact that the phonological categories of intonation, whatever they are, are subject to meaningful gradient variation, or "gradience" for short. The simplest example of this kind of variation involves pitch range. If we emphasize a word by putting a high accent peak on it, we can emphasize it even more by making the high accent peak higher. This kind of thing doesn't happen with *ankles* and *uncles*; phonetic variability in the quality of the first vowel may say something about where the speaker is from, but it doesn't affect the meaning of the words in the way that gradient pitch range modification affects the meaning of an accent.

Importantly, this kind of phonetically gradient variability has the potential to signal pragmatic nuances that seem quite distinct; for example, Hirschberg and Ward (1992) show how a distinction between "uncertainty" and "incredulity" can be signaled by gradient differences in pitch range in the phonetic realization of the English rising-falling-rising contour that they notate L* + H L- H%. It is far from clear how such sharply distinct pragmatic inferences emerge from the application of multiple dimensions of gradient variability to the same intonational category. Nevertheless, as I have argued before (e.g., Ladd 2008a, 151–156), it is important to be aware that this is part of how intonational meaning works. The fact that an intonationally conveyed pragmatic distinction seems categorical does not entail that the corresponding intonational distinction involves categorically distinct phonological elements. Until we understand this better, our phonological analyses are likely to make spurious categorical distinctions.

For example, ToBI draws a categorical distinction between a pitch accent transcribed L + H* and one transcribed H*—informally speaking, a distinction between a rising accent and an accent that is merely high. In the second case, the more or less inevitable rise preceding the peak is considered to be transitional; in the first, it counts as an essential part of the pitch accent's phonetic manifestation. There is little doubt that in some contexts, the presence or absence of a clear accentual rise yields distinct nuances in the pragmatic force of a high accent. Equally, there is little doubt that this distinction is difficult to draw consistently (the evaluation study of intertranscriber reliability in ToBI by Pitrelli, Beckman, and Hirschberg 1994 collapsed the two categories because agreement on the distinction was so poor). This particular descriptive dilemma has a long history that precedes the emergence of AM intonational phonology by several decades. In British school analyses of intonation of the sort pioneered by Palmer (1922), some analysts (including Palmer himself) posited a single category of falling nuclear accent, while others (such as Kingdon 1958 and Halliday 1967) distinguished fall (Halliday's tone 1) from rise-fall (Halliday's tone 5).

In my opinion, the explanation for this situation is that the distinction between L + H* and H* is a matter of meaningful gradient phonetic variability—Bolingerian gradience—within a single phonological category. The extent (and probably the timing) of the rise in pitch preceding the accentual high are manipulated by the speaker to express a variety of nuances, but the basic phonological choice is in all cases "high accent." However, my point here is not to argue for one analysis over the other, but only to point out that the widespread acceptance of ToBI as a standard—together with the fact that the phonetic basis of the distinction is sometimes readily observable—means that many transcribers of English intonation take it as uncontroversial that there is a categorical distinction between L + H* and H*. The specific issue appears settled, and its potential value as evidence for how gradient phonetic variability affects meaning remains unrealized.

For a more complicated case, consider the transcription of what is often called *uptalk*—the use of rising phrase-final contours on statements in a variety of contexts of interaction. There is a substantial set of research findings showing that variation in the pitch level at the beginning of the rise is systematically related to various pragmatic factors such as whether the utterance is intended as a statement or a question (Warren 2016, 33–45 passim) and whether the utterance involves new or given information (Warren 2016, 62–64). These are empirical results, and we want to incorporate them into our understanding. However, because of the problem of gradience, it's not clear how to do this. For a number of investigators who use some version of the ToBI/AM analysis of pitch contours,[1] the obvious way to annotate this difference is to distinguish between an L* and an H* pitch accent, so that uptalk includes both L* H- H% (for statements, for given information, and so on) and H* H- H% (for questions, for new information, and so on). But any such analysis has serious problems.

First of all, it has the problem that the empirical findings are a matter of statistical generalization, not categorical distinction. It is as if we discovered that when speakers intend to talk about the brother of one of their parents, they use *uncle* 87 percent of the time and *ankle* 13 percent of the time, while references to the joint between lower leg and foot involve *ankle* in 79 percent of cases and *uncle* in 21 percent. Obviously, ordinary word meaning doesn't work this way; if intonational meaning does work like this, then we have to adjust our understanding of what a phonological notation system for intonation is supposed to tell us.

Second, in practice, the notational distinction between L* H- H% and H* H- H% results from the fact that transcribers will try to find a way to represent audible differences that convey an intonational nuance, without regard to the phonological implications. The runaway success of ToBI-style tonal transcription systems for intonation has meant that certain categorical notational distinctions (e.g., L* versus H*) are now widely taken for granted as a useful representation of phonetic reality; the notations L* H- H% and H* H- H% are used to express phonetic observations that may actually involve the extremes of a gradient continuum. Yet if we believe (with, e.g., Pierrehumbert and Hirschberg 1990 and Steedman 2014) that the elements of a ToBI string are distinct phonological categories with distinct meanings, then L* H- H% and H* H- H% are simply different strings, like *ankle* and *uncle*, and they should not be regarded as variants of a single basic phenomenon (uptalk) any more than *ankle* and *uncle* can be regarded as variants of the same word. But this is exactly what frequently happens. The theoretical contradiction between strictly interpreting the notations as a phonological analysis and informally lumping the two types together as uptalk, which implies some sort of functional and phonological unity, goes unremarked.

The way to resolve the theoretical contradiction seems obvious to me: we need to recognize fewer phonologically distinct intonational categories, and more dimensions of meaningfully gradient phonetic variation. I'm quite sure that uptalk really is a phenomenon, and we want to be able to describe it in such a way that we can separate uptalk from non-uptalk with the same kind of confidence that we can distinguish *ankle* from *uncle*. But to do this, we cannot also give categorically distinct labels to uptalk-starting-low and uptalk-starting-high; we have to be content to describe those nuances in gradient, statistical, phonetic terms. This is hard to do in a convenient notational system. The choice of one transcription symbol over another to express an audible phonetic difference inclines us to believe that we are dealing with two different phonological categories. In segmental transcription, this belief is often justified, but in intonational transcription, it actively hinders the development of our understanding.

Mark Liberman makes essentially the same point in one of his early *Language Log* posts on uptalk:

> It's also worth mentioning that the form of final rises can vary a lot. The starting point and ending point can move around, with respect to the speaker's pitch range and also with respect to the immediately preceding material. The rate of rising and the alignment with the words of the message also vary. *It remains unclear, in my opinion, which aspects of this variation are phonetic dimensions that speakers can choose to deploy to a greater or lesser degree, like the choice of how fast or loud to talk, and which aspects involve qualitatively different alternatives, like the choice between two different words*. (2006, emphasis added)

Liberman's frank acknowledgment that "it remains unclear" which phonetic features of intonation are gradient and which categorical strikes me as exactly right. The trouble with ToBI is that it doesn't acknowledge the existence of such fundamental uncertainty.

Acknowledgments

I have discussed aspects of the trouble with ToBI in several publications, and in trying to put everything together in this essay, I have made use of various phrases, sentences, and even a few whole paragraphs from those earlier works. Specifically, I acknowledge borrowing from chapter 3 of my book *Intonational Phonology* (2008a) and from my reviews (Ladd 2008b, 2017, respectively) of Jun (2005) and Warren (2016).

Note

1. Strictly speaking, the following discussion concerns a problem not with ToBI proper, but only with the unquestioned extension of ToBI-influenced tonal analyses for use in a wide variety of contexts. I include it because it provides a clear example of the difficulty of dealing with intonational gradience in any symbolic transcription system.

References

Arvaniti, Amalia, and Mary Baltazani. 2005. "Intonational Analysis and Prosodic Annotation of Greek Spoken Corpora." In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by Sun-Ah Jun, 84–117. Oxford: Oxford University Press.

Ashby, Michael, Patricia Ashby, John Baldwin, Frederika Holmes, Jill House, and John Maidment. 1995. "Broad Transcription in Phonetic Training." In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, edited by Kjell Elenius and Peter Branderud, vol. 3, 170–173. Stockholm: Royal Institute of Technology and University of Stockholm.

Beckman, M. E., and G. Ayers Elam. 1997. "Guidelines for ToBI Labelling." *OSU Research Foundation* 3:30.

Beckman, Mary E., Julia Hirschberg, and Stefanie Shattuck-Hufnagel. 2005. "The Original ToBI System and the Evolution of the ToBI Framework. In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by Sun-Ah Jun, 9–54. Oxford: Oxford University Press.

Beckman, Mary E., and Janet B. Pierrehumbert. 1986. "Intonational Structure in English and Japanese." *Phonology Yearbook* 3:255–310.

Birch, Stacy, and Charles Clifton Jr. 1995. "Focus, Accent, and Argument Structure: Effects on Language Comprehension." *Language and Speech* 38:365–391.

Bolinger, Dwight L. 1961. *Generality, Gradience, and the All-or-None*. The Hague, the Netherlands: Mouton.

Brugos, Alejna, Stefanie Shattuck-Hufnagel, and Nanette Veilleux. 2006. "Transcribing Prosodic Structure of Spoken Utterances with ToBI." MIT Open Courseware. https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/.

Cruttenden, Alan. 1994. "Rises in English." In *Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor*, edited by J. Windsor Lewis, 155–173. London: Routledge.

Gibbon, Dafydd. 1976. *Perspectives of Intonation Analysis*. Bern, Switzerland: Lang.

Grabe, Esther. 1998. "Comparative Intonational Phonology: English and German." PhD diss., Max Planck Institute for Psycholinguistics.

Grice, Martine, and Stefan Baumann. 2002. "Deutsche Intonation und GToBI." *Linguistische Berichte* 191:267–298.

Grice, Martine, Matthias Reyelt, Ralf Benzmüller, Jörg Mayer, and Anton Batliner. 1996. "Consistency in Transcription and Labelling of German Intonation with GToBI." Paper presented at the Fourth International Conference on Spoken Language Processing. https://ieeexplore.ieee.org/abstract/document/607958.

Gussenhoven, Carlos. 1984. *On the Grammar and Semantics of Sentence Accents*. Dordrecht, the Netherlands: Foris.

Gussenhoven, Carlos. 1993. "The Dutch Foot and the Chanted Call." *Journal of Linguistics* 29:37–63.

Gussenhoven, Carlos, Toni Rietveld, Joop Kerkhoff, and Jacques Terken. 2003. "Transcription of Dutch Intonation" (interactive course). 2nd ed. http://todi.let.kun.nl/ToDI/home.htm.

Halliday, M. A. K. 1967. *Intonation and Grammar in British English*. The Hague, the Netherlands: Mouton.

Hirschberg, Julia, and Gregory Ward. 1992. "The Influence of Pitch Range, Duration, Amplitude and Spectral Features on the Interpretation of the Rise-Fall-Rise Intonation Contour in English." *Journal of Phonetics* 20:241–251.

Hualde, José, and Pilar Prieto. 2016. “Towards an International Prosodic Alphabet (IPrA).” *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7 (1): 5.

Jun, Sun-Ah, ed. 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Kingdon, Roger. 1958. *The Groundwork of English Intonation*. London: Longman.

Kiparsky, Paul. 2014. “New Perspectives in Historical Linguistics.” In *The Routledge Handbook of Historical Linguistics*, edited by Claire Bowern and Bethwyn Evans, 64–102. London: Routledge.

Ladd, D. Robert. 1978. “Stylized Intonation.” *Language* 54:517–539.

Ladd, D. Robert. 1983. “Phonological Features of Intonational Peaks.” *Language* 59:721–759.

Ladd, D. Robert. 2008a. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Ladd, D. Robert. 2008b. Review of *Prosodic Typology: The Phonology of Intonation and Phrasing*, by Sun-Ah Jun. *Phonology* 25:372–376.

Ladd, D. Robert. 2017. Review of *Uptalk: The Phenomenon of Rising Intonation*, by Paul Warren. *Journal of the International Phonetic Association* 47:369–374.

Leben, William. 1976. “The Tones in English Intonation.” *Linguistic Analysis* 2:69–107.

Liberman, Mark. 2006. “Uptalk Is Not HRT.” *Language Log*. http://itre.cis.upenn.edu/~myl/languagelog/archives/002967.html.

Mayo, Catherine, Matthew Aylett, and D. Robert Ladd. 1997. “Prosodic Transcription of Glasgow English: An Evaluation Study of GlaToBI.” In *Intonation: Theory, Models and Applications: Proceedings of an ESCA Workshop, Athens*, edited by A. Botinis, G. Kouroupetroglou, and G. Carayiannis, 231–234. Athens: ESCA and University of Athens.

Moulton, William G. 1947. “Juncture in Modern Standard German.” *Language* 23:212–226.

Palmer, Harold. 1922. *English Intonation, with Systematic Exercises*. Cambridge: Heffer.

Pierrehumbert, Janet B. 1980. “The Phonology and Phonetics of English Intonation.” PhD diss., MIT.

Pierrehumbert, Janet B., and Julia Hirschberg. 1990. “The Meaning of Intonational Contours in the Interpretation of Discourse.” In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Pike, Kenneth L. 1943. *Phonetics: A Critical Analysis of Phonetic Theory and a Technique for the Practical Description of Sounds*. Ann Arbor: University of Michigan Press.

Pitrelli, John, Mary E. Beckman, and Julia Hirschberg. 1994. “Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework.” Paper presented at the Third International Conference on Spoken Language Processing. https://www.isca-speech.org/archive/icslp_1994/i94_0123.html.

Price, Patti J., Mari Ostendorf, Stefanie Shattuck-Hufnagel, and Cynthia Fong. 1991. “The Use of Prosody in Syntactic Disambiguation.” *Journal of the Acoustical Society of America* 90:2956–2970.

Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. “ToBI: A Standard for Labeling English

Prosody." Paper presented at the Second International Conference on Spoken Language Processing, Banff, Alberta, Canada.

Steedman, Mark. 2014. "The Surface-Compositional Semantics of English Intonation." *Language* 90:2–57.

Venditti, Jennifer. 1997. "Japanese ToBI Labelling Guidelines." *Working Papers in Linguistics, Ohio State University Department of Linguistics* 50:127–162.

Warren, Paul. 2016. *Uptalk: The Phenomenon of Rising Intonation*. Cambridge: Cambridge University Press.

Wightman, Colin, Stefanie Shattuck-Hufnagel, Mari Ostendorf, and Patti Price. 1992. "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries." *Journal of the Acoustical Society of America* 91:1707–1717.

7 The Prosogram Model for Pitch Stylization and Its Applications in Intonation Transcription

Piet Mertens

7.1 Introduction

This chapter describes a generic, integrated approach to the analysis of speech prosody, covering its acoustic, perceptual, and linguistic manifestations, as well as the relationships and mapping among them. Analysis proceeds bottom-up, starting from acoustic parameters, over pitch stylization simulating tonal perception, via the labeling of pitch levels and movements using discrete symbols, so as to ultimately obtain a structured representation of intonation in which pitch events are aligned with prosodic structure, forming prosodic units, with their internally structured pitch contours. Both the initial analysis and its more advanced applications are implemented as a computational system, allowing for a systematic validation of the model and its components.

Prosodic features of speech may be characterized at three major levels of observation: the acoustic, the perceptual, and the linguistic. The last is commonly referred to as the phonological level, because it involves abstraction and reduction of phonetic detail. For completeness, a fourth characterization of prosody should be added, which is the physiological level. Whatever we measure, perceive, or categorize is first produced by the vocal apparatus—by neural instructions to the muscles in the thorax and the larynx. These physiological aspects, however, are not dealt with in this chapter.

Let us briefly sketch the three representations discussed here. First, *acoustic analysis* measures continuous physical parameters of the speech signal, such as fundamental frequency (F0), voicing, intensity, and spectral energy distribution. Spectral information changing over time enables the identification of sounds, syllables, and pauses. Second, perceptual processing of these acoustic events results in an *auditory representation*, which entails a substantial reduction of the variation present in the acoustic signal and produces a segmentation into syllable-sized units. Due to this segmentation, the continuous speech signal is transformed into a sequence of short duration fragments. This process has a crucial impact on the perception of prosodic attributes such as pitch, loudness, and duration. For instance, F0 change is transformed into a sequence of pitch steps and pitch movements. Also, the emergence of syllables gives rise to derived prosodic attributes such as prominence, speech rate, and rhythm. The third observation level, the *phonological characterization*, assumes a minimal set of distinctive prosodic forms (called "tones" in some approaches), associated with either individual syllables or sequences of syllables. The concatenation of such forms results in contours anchored at stressed syllables. These tones and contours are assumed to have a function in speech communication. Although related, the three representations or observation levels differ considerably: one abstract phonological form may correspond at the perceptual level to several forms with slightly different auditory shapes. In addition, the acoustic realization of such forms depends on the phonetic composition of the syllable on which it occurs, it varies with speech rate, and so forth.

When comparing these three representations of prosody for random utterances, the differences among them are striking. Moving bottom-up from acoustics to phonology, one goes from continuous parameters over sequences of simplified shapes to abstract forms noted by discrete symbols. Although different, each of these representations is adequate in some respect and appropriate for some purpose. Acoustic specifications constitute the objective representation of the sound signal, from which all other representations are derived. They are essential in speech synthesis, for instance. But to understand speech, listeners do not need spectrograms or pitch tracks; auditory information suffices. The perceptual representation, as witnessed by handmade transcriptions of prosody, based on auditory information only, shows to what extent perceptual processing transforms the acoustic information, smoothing pitch variation and introducing segmentation, thereby creating a pattern in auditory memory that the listener can describe verbally and reproduce in a graphical form. Finally, the abstract linguistic representation further categorizes these auditory shapes, eliminating free and contextual variation, normalizing pitch range differences between speakers, normalizing temporal aspects such as speech rate and rhythm, generalizing over syllable sequences of various lengths, and introducing structure based on word stress, for example.

A comprehensive theory of speech prosody should aim not only at an adequate acoustic representation of the prosodic parameters or a compact (or even a minimal) characterization of distinctive forms. It should also aim at explaining the mapping between representations of increasing degrees of abstraction. How does the perceived auditory pattern emerge from the acoustic data? Which mechanisms account for the categorization of the auditory shapes into a closed set of basic forms, which we transcribe by discrete symbols? Such conversions are essential, because understanding and simulating these mappings show they are reproducible and predictable. Hence, the (continuous) perceptual representation and the (discrete) linguistic representation are not arbitrary but, rather, have a cognitive status.

The general approach outlined is illustrated in figure 7.1. Its three panels broadly match the acoustic, perceptual, and linguistic observation levels discussed. The top panel shows the acoustic level, illustrated by the F0 trace. The middle panel, corresponding to the perceptual level, shows the stylized pitch curve aligned with the syllable chain, together with other parameters used in the stylization (e.g., intensity, voicing, phoneme and syllable boundaries, and pauses). Finally, the bottom panel shows a discrete transcription using symbols for pitch levels (low, mid, high, bottom, and top) and pitch movements (rise, fall, level). For instance, the symbol HF, for “high fall,” indicates that the corresponding syllable contains a large downward pitch movement starting from a high pitch level.

Every symbolic transcription of intonation makes theoretical assumptions about its underlying structure: its basic elements and their possible combinations, their function in language, and their interaction with the units and syntactic organization of the segmental layer. The impact of such theoretical assumptions is demonstrated by the bewildering number of competing intonation models that have been proposed. The distinctive feature of the approach presented here is that it proposes a comprehensive prosody model, combining the acoustic, perceptual, and discrete symbolic specifications, as well as the mapping between successive representation levels. Most importantly, the perceptual and symbolic representations of figure 7.1 were computed automatically from the speech signal and the phonetic alignment, strongly supporting the claim that

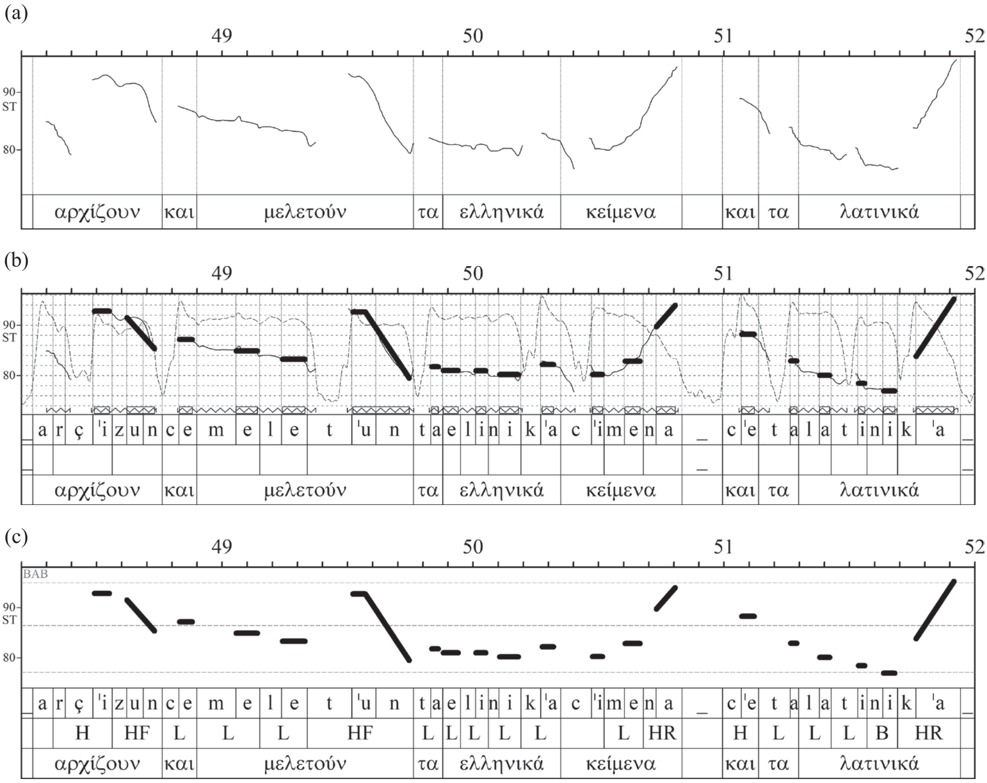


Figure 7.1
Acoustic, perceptual, and symbolic representations of the pitch contour of the Greek utterance “αρχίζουν και μελετούν τα ελληνικά κείμενα και τα λατινικά” (They start studying the Greek texts and the Latin ones), pronounced by a male speaker. The upper panel shows fundamental frequency on a semitone scale. The bold line in the center panel shows the stylized pitch based on a model simulating tonal perception. The central panel also includes acoustic parameters used by the stylization, such as intensity (dashed line) and voicing (sawtooth line), as well as the segmentation into sonority peaks (shown as boxes). The three annotation layers (or “tiers”) in the central panel indicate, from top to bottom, phonetic symbols (including stress marks), syllables, and words. The bottom panel shows a symbolic transcription of pitch levels and movements in the second tier (aligned with the syllables). The detected pitch range is visualized by the three horizontal dashed lines, indicating top and bottom of the pitch span and the median pitch.

by simulating auditory perception and, in particular, tonal perception, the acoustic signal may be mapped onto its perceptual representation, which in turn may be categorized into discrete symbols. The latter are compatible with those posited in linguistic models lacking a computational model. The mapping in the opposite direction, from discrete symbols to the acoustic speech signal, has already been illustrated in text-to-speech synthesis (Mertens et al. 2001).

The remainder of this chapter is organized as follows. Section 7.2 deals with pitch contour stylization simulating pitch perception. After describing the pitch perception in speech (section 7.2.1), it outlines the Prosogram algorithm for simulating it (section 7.2.2) and discusses the validation of this model (section 7.2.3). Section 7.3

summarizes the prosodic attributes that are computed within this computer program for each syllable, as well as for sequences of syllables. One such attribute, pitch range, is discussed in more detail. These data may be used in experimental research and speech applications. The next two sections present two applications of the Prosogram model in the area of prosody transcription: in section 7.4, the automatic transcription of pitch levels and movements observed at the level of the syllable, and in section 7.5, fully fledged intonation transcription, making explicit underlying structure, prosodic units, and pitch contours. The final section (7.6) discusses the major assumptions underpinning this approach.

Before we set out, a quick reminder on pitch scales. In voiced speech, the *fundamental frequency, F0*, of the quasi-periodic waveform corresponds to the frequency of vocal fold vibration. It is an acoustic property. *Pitch* is the perceptual correlate of F0. Whereas frequency is measured in hertz, various scales are used for pitch measurement (Nolan 2003): hertz, semitone (ST), Mel, Bark, equivalent regular bandwidth (Hermes and Van Gestel 1991), and octave-median (De Looze and Hirst 2014). The musical scale in ST indicates a *pitch interval*, that is, the distance in pitch between two points in time (for instance, between two musical notes, two sounds, or the start and end of a pitch glide), rather than an absolute pitch or frequency. The distance *D* in ST between two frequencies f_1 and f_2 is given by: $D = 12 \log_2 (f_2/f_1)$ (Baken and Orlikoff 2000, 148). When f_2 is twice as high as f_1, they are separated by one octave. The ST scale divides the octave into twelve equal steps on a logarithmic scale. When speakers with a different vocal range (for instance, a man and a woman) pronounce the same utterance while imitating its intonation, their fundamental frequency range will probably differ, although the pitch intervals (in ST) used in their intonation contours may be similar. The ST scale thus allows for pitch contour normalization, maintaining the actual pitch intervals independently of absolute pitch. The ST scale may also be used as an absolute scale, by setting f_1 in the formula for *D* to a fixed reference frequency. (The illustrations in this chapter use ST relative to 1 Hz. The reference frequency may be selected arbitrarily: a given pitch interval remains the same size in ST, whether the reference frequency is set to 1 Hz or 100 Hz. For more details on the ST scale, see Mertens (2020, section 9.1).

7.2 Pitch Stylization Based on a Tonal Perception Model

Stylization is a process that simplifies the F0 curve of the speech signal. It removes local F0 variations which are perceptually irrelevant, because they are too small or too short to be perceived, because they occur in spectrally unstable regions, such as the transition between consonants and vowels (micro-prosody or micro-intonation), and so forth. By doing so, it maintains only those aspects that potentially have a function in speech communication. The type of stylization described here aims at simulating the auditory representation as it emerges from perceptual processing. In this way, it helps to reveal the underlying intonation pattern.

In general, pitch stylization makes the assumption that pitch contours of utterances can be adequately synthesized and hence be represented by a sequence of basic shapes, such as straight lines or curves. In his detailed account of stylization, Hermes (2006) distinguishes two major approaches, depending on the nature of the elementary parts. The first approach obtains a sequence of elementary shapes, either straight lines ('t Hart, Collier, and Cohen 1990; Scheffers 1988; Spaai et al. 1993) or curves (Taylor 1994, 2000; Hirst, Nicolas, and Espesser 1991; Hirst et al. 2000). In this approach, the turning points of the F0 contour determine the boundaries between elementary shapes.

In the second approach, the stylization takes the form of a sequence of syllabic tones, that is, simple or compound pitch movements (including steady pitch) associated with syllables (d'Alessandro and Mertens 1995; Mertens 2004a, 2004b). Obviously, syllabic tones are also elementary shapes. But whereas in the first approach, the pitch contour is seen as a continuum and the boundaries of elementary units are determined by pitch change only, the second approach acknowledges the impact of spectral change on tonal perception. As a result, the segmentation based on pitch change is preceded by a segmentation based on spectral change.

A closer look at the first major approach shows that some of these models implicitly take into account spectral change and amplitude variation. This is illustrated by the IPO approach ('t Hart, Collier, and Cohen 1990) (IPO refers to the research institute "Instituut voor Perceptie Onderzoek" in Eindhoven, the Netherlands, where this approach was developed), which distinguishes early and late pitch movements depending on the temporal alignment of the movement with vowel onset and vowel end (Hermes 1987). But other models, such as the quadratic spline modeling (Hirst, Nicolas, and Espesser 1991), claim that the perception of pitch variation in speech is continuous (Hirst 2011) and reject any form of segmentation. In such a view, turning points may occur anywhere in the signal, within vowels or consonants, or even in silent portions.

The remainder of this section takes a closer look at pitch stylization based on tonal perception. It lists various perceptual phenomena observed in psychoacoustics, which are simulated in the stylization model, as well as in the Prosogram tool, which is an implementation of this model using the Praat speech analysis program (Boersma and Weenink 2012).

7.2.1 Tonal Perception

7.2.1.1 Spectral instability and its impact on pitch perception Spectral change and amplitude drop at the transition between sounds (consonants and vowels) can perceptually mask F0 change during this transition ('t Hart, Collier, and Cohen 1990, 36). Experiments by House (1990) show that an F0 variation—for instance, a linear rise of fixed duration and size—is perceived differently, as either a pitch movement or an abrupt change of pitch, depending on its location relative to the vowel onset in the syllable. This phenomenon (known as the *spectral stability hypothesis*) is attributed to spectral change at the transition between sounds, which lowers sensitivity to pitch change. The amount of spectral change at sound transitions depends on the nature of the sounds involved; nasals and liquids are more similar to vowels than fricatives, for instance. A more detailed segmentation model based on spectral change also takes into account the nature of the sounds in the syllable rhyme (House 1995, 1996). This perceptual phenomenon explains why hearers are able to distinguish pitch movements occurring between adjacent syllables from those occurring within one and the same syllable.

7.2.1.2 Intrasyllabic pitch movement and the glissando threshold Some syllables are perceived with a stable pitch, others with a changing pitch. In music, the latter type would be called a *glissando*. For a gliding pitch, the observed F0 change will exceed a threshold, which depends on the pitch interval covered by the glide and its duration ('t Hart, Collier, and Cohen 1990). Listening experiments with natural or synthetic speech (Rossi, 1971; Rossi et al. 1981; Mertens, Beaugendre, and d'Alessandro 1997) show that many intrasyllabic F0 variations observed in speech are indeed perceived

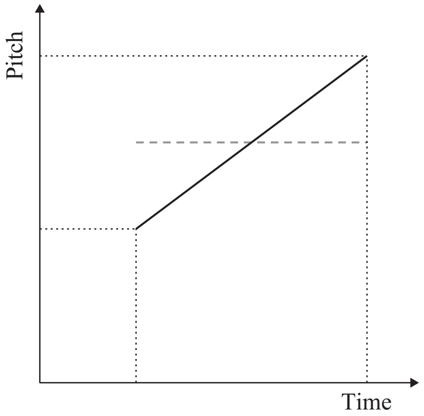


Figure 7.2
Effect of the GT for an F0 variation (solid line) with a given duration and size (dotted lines). When the combination of duration and size is lower than the GT, the F0 variation is perceived as a sound with stable pitch (dashed line).

as glissandi. To decide whether a syllable is perceived with a stable or a gliding pitch, the combination of duration and size of the corresponding F0 variation is compared to the glissando threshold (GT), as illustrated for a uniform F0 change in figure 7.2. When they exceed the GT, a gliding pitch is obtained (solid line), otherwise a stable pitch is perceived with a frequency approaching the median F0 of the variation (dashed line).

The GT was first observed and measured in experimental settings, using stimuli with known properties, such as pure tones or synthetic vowels, presented repeatedly, separated by silence, hence followed by a pause. Natural, continuous speech is characterized instead by spectral instability due to sound concatenation and by the occasional presence of pauses. As a result, the threshold is higher in connected speech than for isolated sounds. House (1995) shows that the presence of the pause facilitates pitch perception—in other words, it lowers the GT. (For a detailed description of the measurement of the GT, see d'Alessandro and Mertens 1995, section 2.5.2.)

7.2.1.3 Simple and compound pitch shapes Hearers have the ability to discriminate between simple and compound pitch movements occurring within a syllable. For compound movements, the successive parts, or tonal segments, may have slopes of opposite direction (e.g., a rise followed by a fall) or of same direction but different rate (e.g., slow rise followed by fast rise, or a level part followed by rise). Whether a pitch variation is perceived as two or more successive pitch movements rather than as a single movement depends on the amount of slope change between the successive tonal segments (see 't Hart, Collier, and Cohen 1990). In figure 7.3 the solid line shows a compound pitch variation consisting of two uniform parts of different slope. When the slope change is sufficiently large, the pitch variation is perceived as a sequence of two parts of different slope, otherwise as one part with a uniform slope, corresponding to the dashed line. In d'Alessandro and Mertens (1995) and in the Prosogram tool, this phenomenon is modeled by the differential glissando threshold (DGT), expressed as $g_2 - g_1$, with g_1 and g_2 the slopes in ST/s of the first and the second tonal segment. (For the DGT measure, see d'Alessandro and Mertens 1995, section 2.5.3.)

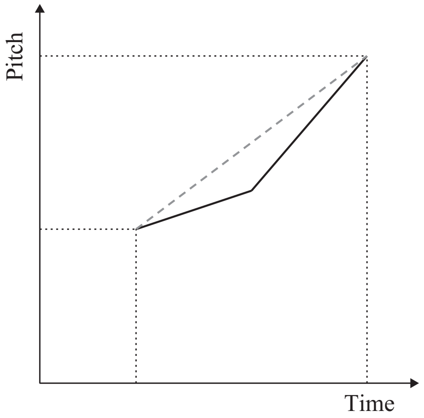


Figure 7.3
Effect of the DGT for a compound F0 variation (solid line) with a given duration and size (dotted lines). When the slope change exceeds the DGT, the variation is perceived as a sequence of two parts with a different slope, otherwise as a single rising pitch movement with constant slope (dashed line).

7.2.1.4 Minimal duration of tonal segments In addition to the two thresholds just mentioned, there appears to be an additional constraint on the minimal duration of a tonal segment. For compound pitch movements, consisting of two or more parts of uniform slope and with sufficiently large slope changes between them (i.e., exceeding the DGT), each segment needs to have a minimal duration of about 0.035 s. A shorter segment is not perceived as an individual tonal segment but is perceptually merged with the adjacent segment that is most similar.

7.2.1.5 Temporal alignment of pitch gestures with speech sounds Pitch events that are perceptually similar vary considerably in their acoustic shape, depending on the nature of the sounds over which they occur. We are not referring here to micro-prosody (also called micro-intonation), which among other things includes F0 variation due to sound nature (e.g., vowel aperture or consonant voicing, so-called *intrinsic micro-prosody*) or phonetic context (e.g., the pitch perturbation during vowel onset after voiceless obstruents, so-called *co-intrinsic micro-prosody*). Rather, we refer to the way in which a pitch event (a target or a movement) is temporally aligned with the segmental layer. Consider a high pitch target on a vowel, following a low pitch target on the preceding syllable. The transition from low to high may show up as either a pitch step (a discontinuity) or a pitch rise during syllable onset, depending on the voicing of the consonants in the onset. When the syllable onset contains highly sonorous sounds, such as a nasal or a glide, this usually results in a slower rise, spread over the onset and the vowel. So, a given pitch target sequence results in quite different F0 contours in the syllable onset, depending on phonetic context. Except for trained phoneticians, listeners generally are not aware of such acoustic differences. Obviously, similar observations can be made for the syllable rhyme, where the pitch movement will be spread over the sonorous segments in the coda when present.

Although the impact of intensity and spectral change on the perception of pitch events was acknowledged in earlier research (e.g., Rossi 1978), such phenomena have largely been ignored in dominant analysis frameworks since the 1980s. The IPO approach ('t Hart, Collier, and Cohen 1990) reduces intonation to a concatenation of standardized linear pitch movements (in the log F0 versus time domain) anchored at a syllable (either early, late, or very late pitch movements) or a prosodic boundary. The variation due to phonetic context was considered perceptually irrelevant. Conversely, in the autosegmental framework (for an overview, see Grice 2006 or Ladd 2008), intonation is characterized as a sequence of tones (or pitch levels) associated with particular syllables (called "tone-bearing units"), and the pitch of other speech portions is supposed to be accounted for by interpolation between these targets. Recently, however, segmental context has gained renewed attention in the context of the tonal center of gravity model (Barnes et al. 2014) of the perceptual integration of pitch change, weighted by segment sonority.

Given the acoustic variability of perceptually equivalent pitch contours, a prosody model should explicitly take into account the temporal alignment of pitch events with the segmental layer.

7.2.2 The Stylization Algorithm

This section describes how the above-enumerated perceptual phenomena are simulated in the stylization algorithm.

7.2.2.1 Segmentation into syllabic nuclei or rhymes The *syllable* is a central unit for the characterization of prosodic attributes such as prominence, stress, syllable length, speech rate, rhythm, and pitch movements. While some pitch events are synchronized with individual syllables (in particular, the stressed syllable), others are carried by a sequence of contiguous unstressed syllables ('t Hart 1998). Even for such gradual movements, the boundaries of the movement coincide with those of a syllable. Hence, the analysis of prosody requires identifying syllable boundaries.

The syllable, however, is a phonological unit. From an acoustic perspective, the exact location of syllable boundaries is often unclear. Whereas plosives and voiceless fricatives are characterized by abrupt changes in the acoustic signal, boundaries are acoustically less clear for glides, nasals, laterals, approximants, and diphthongs. Also, from a production perspective, sounds may be ambisyllabic: in this case, the closure of a consonant is part of a syllable's coda, whereas its release starts the next syllable's onset. For pitch perception, what matters is not the boundaries of the phonological syllable but rather those of the central part of a syllable, characterized by high sonority and relative spectral stability. This central part will be referred to as the *syllabic nucleus*. (Note that this term is already used in phonology to designate the vowel part of a syllable.)

In addition, fundamental frequency is defined only for voiced sounds, and intensity and spectral distribution may vary considerably during the syllable. These changes in periodicity, intensity, and spectrum affect the perception of pitch changes (Rossi 1978).

The procedure for syllabic nucleus detection adopted by Prosogram simulates perceptual properties while using acoustic parameters. The unit is identified as the part of the voiced portion of a syllable, located around the intensity peak within the vowel, where intensity decrease is below a given threshold. For the left boundary, this threshold is at 2 dB below the peak. For most recording conditions, this provides an acceptable approximation of the vowel onset. For the right boundary, the threshold is chosen so as to include the voiced, high-intensity part of the rhyme. More precisely, the intensity drop threshold is equal to 80 percent of intensity difference between the peak and the local minimum in the time interval between the peak and syllable end, with a minimum of 3 dB. When the resulting time interval contains an abrupt pitch change, most often resulting from an octave jump pitch detection error, the interval is truncated at this point to avoid erroneous pitch values. A minimal duration (0.025 s) for the resulting syllabic nucleus is also required.

For best accuracy, the segmentation may be guided by the phonetic and syllabic alignment provided by the corpus annotation. In this case, these alignments are used to locate the time window where the syllabic nucleus is to be found, more precisely to locate the vowel and the syllable rhyme, consisting of the vowel and the coda. The resulting segmentation is illustrated in figure 7.4.

In addition, the Prosogram tool provides alternative segmentation types, which may be selected depending on the temporal alignments available for the corpus: phonemes only, syllables only, or vowels only.

For corpora without phonetic alignment, automatic segmentation is available, which is based on intensity of band-pass filtered signal (300–3,500 hertz), intensity, voicing, and F0 discontinuities. This is illustrated in figure 7.5.

7.2.2.2 Pause detection Silent pauses play a role in discourse—for instance, by marking the boundaries between major syntactic constituents or in speaker turn management. Silent pauses also affect the perception of pitch events by lowering the GT (House 1995).

Pause detection may be complicated by background noise and speech dynamics, which depend on recording conditions and vary from one corpus to the next. Most pause detection algorithms exploit the difference in intensity between voiced speech and silent pauses. In case of background noise, the intensity of nonspeech increases. In case of changing recording conditions, such as gain adjustments during recording, including automatic gain control, it is difficult to select an optimal threshold.

As a practical solution, the following procedure is used. The segmentation into syllables results in a sequence of syllabic nuclei. The gap between the end of one nucleus and the beginning of the next provides a rough estimate of pause length. When this gap exceeds 0.35 s, it is interpreted as a pause. In the illustrations, detected pauses are marked by a *P* at the start time of the gap.

7.2.2.3 Pitch stylization Stylization is applied to the F0 data of each syllabic nucleus obtained by the segmentation (see section 7.2.2.1). The algorithm, which is described in detail in d'Alessandro and Mertens (1995), may be summarized as follows. For each syllabic nucleus, the pitch contour is divided into one or more parts of uniform slope (either level, rising or falling) called *tonal segments*. First, turning points in the pitch contour are located by order of importance. For a given time interval, a candidate turning point is found at the time of maximum distance between the observed F0 and a straight line connecting the F0 values at the start and end of the interval under consideration. This candidate turning point is accepted only when all following conditions are met: (i) the F0 distance at the turning point exceeds 1 ST, (ii) both resulting segments exceed the minimal duration of a tonal segment, (iii) at least one of the segments is a glissando, and (iv) the slope change at the turning point exceeds the DGT. The same procedure is applied recursively to the time intervals on both sides of a valid turning point. Recursion halts when no additional turning point is found. After the slicing into tonal segments, the perceived pitch at the turning points is estimated. If the pitch variation of a tonal segment is below the GT, it is replaced by a level slope, with a pitch equal to the median F0 in the corresponding time interval.

In the current model, perceptual thresholds are dynamically adjusted for continuous speech. For a syllable preceding a pause, the GT is set to the value for isolated

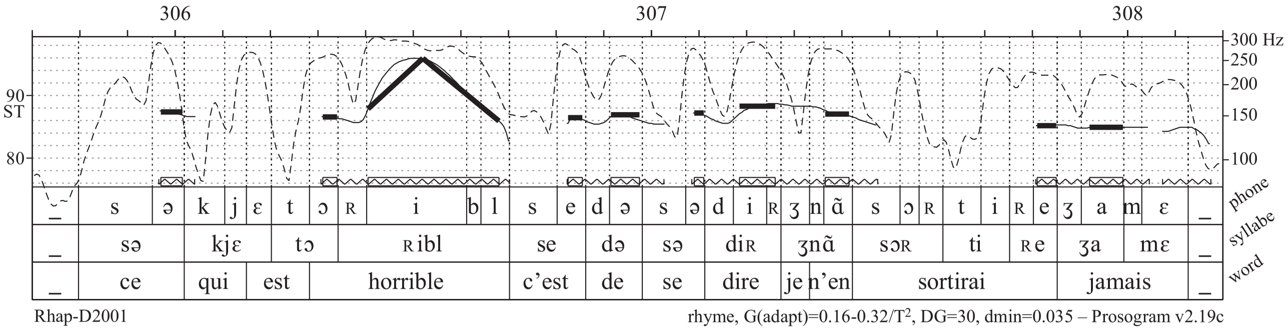


Figure 7.4

Illustration of segmentation into syllabic nuclei guided by phonetic alignment, for the French utterance, "Ce qui est horrible c'est de se dire je n'en sortirai jamais" (What is horrible is to say to oneself: I will never get out of this situation), taken from the Rhapsodie corpus (recording Rhap-D2001; Lacheret-Dujour et al. 2019). The upper part shows the acoustic parameters of intensity (the dashed line), voicing (sawtooth), and fundamental frequency (thin line), mostly covered by the thick line of the stylized pitch. Pitch is plotted on an ST scale (relative to 1 Hz), with horizontal calibration lines at two ST steps. The lower part shows various corpus annotation tiers: phonetic alignment, syllables, and words. The syllabic nuclei appear as small boxes on top of the sawtooth voicing line. For the syllable [ʀibl], the nucleus includes the vowel and part of the coda. The syllables [kjɛ], [sɔʀ], and [ti] are analyzed as unvoiced; consequently no nucleus is detected.

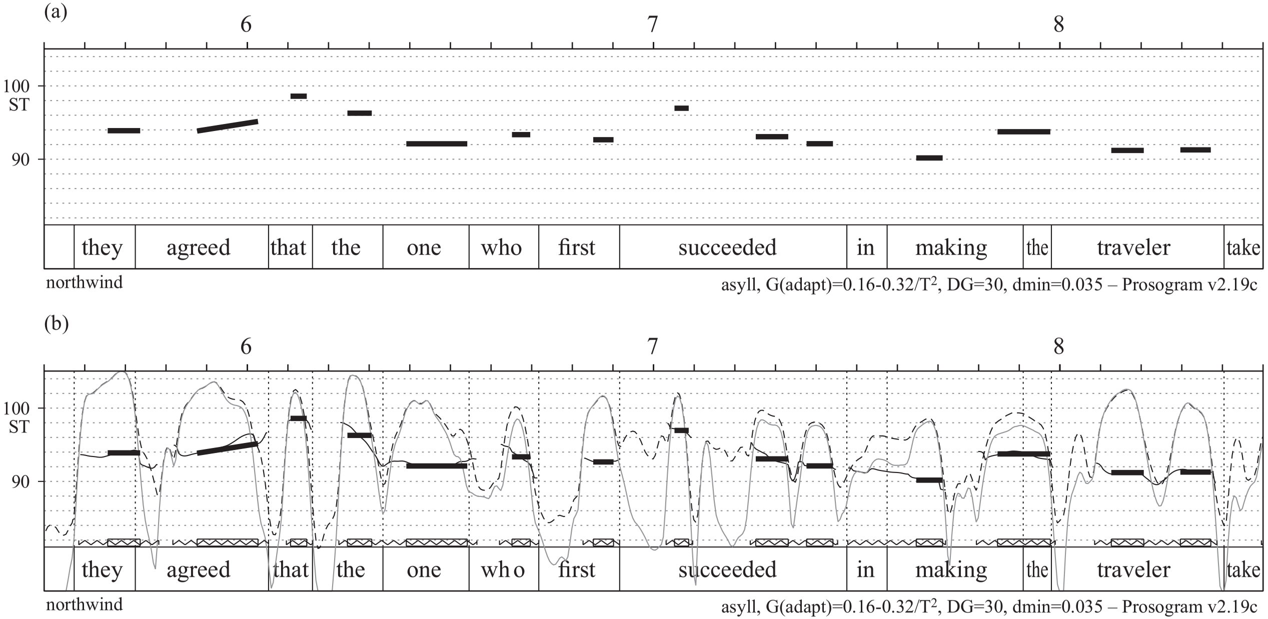


Figure 7.5
Illustration of the automatic segmentation into syllabic nuclei based on acoustic information only, for the utterance, “They agreed that the one who first succeeded in making the traveler take . . . ,” from “The North Wind and the Sun” passage in American English (International Phonetics Association). The upper panel shows the stylized pitch of syllabic nuclei and the word annotation (not used in the segmentation process). The lower panel also shows the acoustic parameters of intensity (dashed line), intensity of the band-pass filtered signal (continuous solid line with larger dips), voicing (sawtooth), and F0 (thin line interrupted at unvoiced portions, mostly covered by the bold line of the stylization). Pitch is plotted on an ST scale, with horizontal calibration lines at two ST steps.

sounds ($GT = 0.16/T^2$), whereas elsewhere, it is twice as high ($GT = 0.32/T^2$), resulting in lesser sensitivity to pitch glide. The DGT is set to 30 ST/s, and the minimum duration of a tonal segment to 0.035 s.

The stylization results in a representation of the audible pitch events in an utterance, which is less complex than the acoustic data. The obtained stylization is shown as a bold line in all figures in this chapter.

7.2.2.4 Differences between implementations of the model The tonal perception model of d'Alessandro and Mertens (1995, section 2.4) includes a short-term integration of F0, which models the limitation of the auditory system to follow rapid changes in pitch. This phenomenon was observed in a study on the perception of vibrato by d'Alessandro and Castellengo (1994), who estimate the perceived pitch using a time-windowed average pitch model. This F0 integration is not incorporated in the Prosogram, because it appears that for very steep F0 falls, pitch intervals and slopes exceed those of vibrato. However, Prosogram uses another type of F0 integration. For F0 variations below the GT, the perceived pitch is approximated by the median F0.

The Prosogram implementation provides more segmentation types than the model of 1995, enabling a more accurate location of the syllabic nucleus when phoneme and syllable alignment are available. In addition, the perceptual thresholds may be adjusted for continuous speech while being sensitive to pauses (see section 7.2.2.3).

7.2.3 Validation through Resynthesis

The stylization based on tonal perception was tested in an experiment (d'Alessandro and Mertens 1995) using TD pitch synchronous overlap and TD-PSOLA (Time Domain Pitch Synchronous Overlap-and-Add) resynthesized versions of sentences, using either the original F0 contour (stimulus V1) or the stylized F0 contour, for three threshold configurations (stimuli V2, V3, V4). The threshold values were chosen such that the stylizations ranged from narrow (V2), to intermediate (V3), to coarse (V4). The thresholds of V2 are those observed in experiments using isolated pure tones ($GT = 0.16/T^2$ and $DGT = 20$ ST/s). In V3, the GT is twice that of V2. V4 uses very high thresholds ($GT = 0.64/T^2$ and $DGT = 60$ ST/s). The subjects listened to sentence pairs, containing one stimulus with the original F0 contour (V1) and one with a stylized contour (V2, V3, V4), in randomized order. They had to judge whether the sentences were identical, so the same-different paradigm is used. For untrained listeners, it is very difficult to judge intonation independently from other aspects of speech. The percentage of correct responses is expected to be very high for stimulus pairs V1-V1 and V1-V4 because the former pair is identical and the latter is clearly different as a result of coarse stylization. For V1-V2, the score is expected to be close to that for V1-V1, and for V1-V3, the score is expected to be slightly lower than that for V1-V2. These hypotheses were confirmed by the results. A stylized contour using thresholds observed for isolated sounds (V2 stimuli) is hardly distinguishable from the original. For other threshold settings, the proportion of items judged different correlates with threshold magnitude. Still, the impact of threshold magnitude is moderate and smaller than the role played by syllable segmentation. There appear to be rather important differences between the scores of individual subjects. Also, some part of the error is explained by stimulus differences introduced by resynthesis itself.

Mertens, Beaugendre, and d'Alessandro (1997) report a similar experiment comparing two stylization approaches: the automatic stylization based on tonal perception (ATS) and the manual close-copy stylization (CC), in which the F0 data is manually approximated by the standardized pitch movements of the IPO model ('t Hart, Collier, and Cohen 1990). Stimuli were prepared using Linear Predictive Coding (LPC) synthesis. The subjects judged whether a stimulus pair was identical or not. A pair consisted of either two stimuli with the original F0 contour (V1) but resynthesized, or one V1 stimulus and one stimulus with the stylized contour obtained with either the ATS (V2A) or the CC (V2B) model. The results indicate that 93 percent of the V1-V1 pairs are judged identical (although 100 percent were identical), whereas V1-V2A and V1-V2B give, respectively, 90 percent and 88 percent identical responses. This shows that resynthesized stimuli of both models are very close, and on average, the ATS model performed slightly better than the CC model did. This may be explained by the fact that the ATS stylization results in a larger number of line segments.

7.3 Prosodic Features Obtained during Stylization

7.3.1 Prosodic Features of Individual Syllables

During stylization, a wealth of prosodic features are computed for each syllable (more precisely for the syllabic nucleus). They are available for statistical analysis (as illustrated by the *prosodic profile*; see Mertens 2020, section 5.3) or further processing, as discussed in sections 7.4 and 7.5. Pitch-related features for each syllable include start, end, maximum, minimum, mean, and median pitch; cumulated upward pitch intervals; cumulated downward intervals; pitch trajectory (sum of absolute upward and downward movements); and pitch interval between the end of the preceding syllable and the start of the current one. All pitch intervals are expressed in ST. Pitch values normalized for the global pitch range of the individual speaker are also available (see section 7.3.2.1). Another feature is the dynamic nature of the syllable (whether it contains a glissando). Duration measures include nucleus duration, as well as vowel, syllable, and rhyme duration, for corpora including phonetic and syllable alignment. Intensity-related measures are peak intensity (measured in dB) and peak loudness (measured from the excitation spectrum). Pause length is available too (see section 7.3.2.2).

A second set of derived, context-related prosodic features includes feature prominence values. A syllable is prominent for a given feature such as duration, loudness, or pitch, when it stands out from the adjacent syllables. This is quantified as the distance between the syllable's value for feature *f* (e.g., duration) and the mean of the corresponding values in the context. The latter typically consists of two preceding syllables and one following syllable but may be selected otherwise. The context is restricted to a time window (0.5 s on either side of the target syllable) and bounded by the presence of a pause. (The distance is expressed as a proportion or a difference, depending on the linear or logarithmic scale of the feature values.)

7.3.2 Prosodic Features for Sequences of Syllables

Some speakers use a narrow pitch range, others a wide one. Pitch movements may be frequent or rare. Speech may be fast or slow. The prosody of an individual speaker, a speaking style, or a speech fragment is characterized by global prosodic attributes, such as pitch range, speech rate, melodic variability, and isochrony. These attributes are easily derived from the primary prosodic features already mentioned.

7.3.2.1 Pitch range and its normalization The term *pitch range* is often used ambiguously, indicating either the vocal range (tessitura, global pitch span) of a speaker, the register (high, modal, low) used by a speaker (Rietveld and Vermillion 2003), the local pitch span of a stretch of speech, or even the excursion size of a single pitch movement (Gussenhoven 2004). The pitch range of a speaker may be locally reduced or widened; its central value may be locally raised or lowered (as for register change).

Both global and local pitch range may be characterized by two variables (Ladd 1996, 260–261; 2008, 198): overall level and span. *Overall level* (or *key*, in the terminology of Hirst 2011, 71) is often approximated by the mean or median F0 of the speaker. The span designates the F0 range used by the speaker, to a first approximation the distance between maximum and minimum pitch. Some speakers use a wide span, others a narrow one. The span may be expressed in various ways, but it is usually specified in ST, that is, on a logarithmic scale. As Ladd (2008, 198) observes, overall level and span are often conflated because changes of key and changes of span may be hard to distinguish: raising a high tone produces the same effect as widening the span.

Ladd (1996, 267–269; 2008) examines several quantitative models of pitch range. According to Rietveld and Vermillion (2003), the F0 of low-pitch targets forms the strongest cue to perceived register (better than mean pitch, median pitch, or mid value), and the pitch interval (in ST) between low and high targets provides an estimate of perceived register width. De Looze and Hirst (2010) measure pitch range using the pitch targets obtained by a Momel stylization and propose an algorithm to detect changes in local pitch range.

Within the Prosogram tool, the global pitch range of a speaker is estimated on the basis of the distribution of representative pitch values in the corpus. To exclude pitch detection errors occurring at the transition between a voiceless consonant and a vowel or a glide, pitch data is limited to the sonorant part of the syllable, obtained using the segmentation into syllabic nuclei (section 7.2.2.1). Moreover, pitch discontinuities, often related to octave jumps, are located, and among the parts with continuous pitch, the one nearest to the median F0 of the speaker is selected. Finally, clear outliers (more 18 ST off the median F0 of the speaker) are discarded. To deal with intrasyllabic pitch movements appropriately, two pitch values per syllable are used: the maximum and minimum pitch, after stylization. (For syllables with level pitch, these two values are identical.) The bottom and top of the global pitch range are identified as the second and ninety-eighth percentiles of this distribution.

To compare intonation contours pronounced by different speakers, independently of their pitch range, a normalization of pitch values may be obtained. Normalized pitch contours not only show a high degree of interspeaker agreement, but also greater similarity between pitch contours pronounced by the same speaker in different paralinguistic conditions. Various types of normalization have been proposed, such as the *z*-score of F0, in either hertz or semitones (Jassem and Kudela-Dobrogowska 1980; Poiré and Kaminskaia 2004; de Moraes and Rilliard 2014). Bardiaux and Mertens (2014) instead normalize pitch contours by scaling the stylized pitch values (in ST) by the speaker's pitch range (measured as already described), such that the bottom and top correspond to 0 and 100 percent of the pitch range, respectively.

7.3.2.2 Other prosodic attributes Prosodic features obtained using Prosogram may be used to analyze and detect prosodic attributes and events such as prominence (Goldman, Avanzi, et al. 2007; Christodoulides and Avanzi 2014), prosodic boundaries (Mertens and Simon 2013; Christodoulides and Simon 2015), and turn-taking cues (Heldner and Włodarczak 2015) or to characterize speech styles (Goldman, Auchlin, et al. 2007). They have been used in various other domains, including pathological speech and music research (the relation between speech prosody and music; Patel, Iversen, and Rosenberg 2006). Because this chapter focuses on pitch stylization and intonation transcription, these aspects will not be discussed here.

7.4 Automatic Symbolic Transcription of Pitch Level and Movement

The pitch stylization described in the previous sections illustrates the mapping between the acoustic and perceptual manifestations of prosody and provides a representation of auditory pitch events, which, although associated with individual syllables, is still continuous in nature. Linguistic analyses of prosody, in contrast, use discrete elements to represent distinctive pitch contrasts in speech. Depending on the linguistic model, these elements will be pitch levels, pitch movements (seen as primitives), tones, or "pitch accents" (Bolinger 1958). They are associated with particular positions in the speech chain, such as the syllable carrying word stress, the boundaries of a syllable sequence, or a boundary of some kind. Given our bottom-up perspective, the question arises whether it is possible to obtain such a discrete linguistic representation starting from the auditory representation.

Most linguistic analyses of prosody include the notion of stress or a related attribute, be it word stress, sentence stress, nuclear tone, pitch accent, or prominence. This prosodic attribute of stress is not available in the stylization described so far. If stress is marked by acoustic features, then it should be possible to detect it using such features. If stress is ultimately a structural property (Gussenhoven 2004), then stress location may be obtained from other sources such as lexical information and syntactic structure. The latter approach is discussed in detail in section 7.5.

Independent of the presence of stress, the categorization of syllabic pitch contours may still be based on principles common to many intonation languages. This section discusses the nature of such principles and proposes a system to categorize pitch contours into discrete elements, as described in Mertens (2013, 2014, 2019).

The ambitious goal of automatic transcription of prosody according to a linguistic annotation scheme may be subdivided into two major parts. The first part provides a generic transcription of pitch levels and movements associated with individual syllables. The second part identifies positions in prosodic structure, mainly stress and boundaries. Both aspects need to be combined according to the prosodic grammar of the language in question, by interpreting pitch levels and movements in terms of these structural positions.

7.4.1 A Generic Transcription of Pitch Levels and Movements

Pitch variation is produced in the larynx; the biophysical properties of the larynx and vocal folds constrain the pitch variations that may occur in speech: maximum rate of pitch change, excursion size, and so forth (Xu 2005). The possible melodic distinctions are also constrained by the limited resolving power of the auditory system ('t Hart, Collier, and Cohen 1990, 74). For speakers of intonation languages (without hearing impairment), the auditory capabilities are fairly similar. So from the perspectives of speech production and perception, there are no fundamental differences between intonation languages, as far as possible pitch movements are concerned; the constraints on pitch variation are basically the same. It seems feasible then to look for a generic transcription.

Languages will differ in the way pitch variations are combined into larger structures, such as pitch contours, or associated with other prosodic or segmental features, such as lexical stress and phrase boundaries. But in the end, every contour is a concatenation of pitch events occurring on successive syllables. The question then becomes how such contours should be characterized.

Ladd (2008) distinguishes two major approaches to pitch contour characterization, a distinction reminiscent of the levels versus configurations debate (Bolinger 1951). The first view describes pitch contours as configurations, that is, as a sequence of components defined relative to what precedes. Consider a contour consisting of low unstressed syllables, followed by a large high fall on the stressed syllable. Such a configuration specifies the location (stressed syllable, unstressed syllable, boundary), the direction of a movement (rise, fall, level), its melodic interval (large, small), as well as the pitch levels resulting from such intervals (high, low). This approach assumes that a given configuration can occur at various places in the speaker's pitch range without affecting contour identity or function. The second view factors out pitch variation related to the individual speaker or paralinguistic aspects (e.g., bored speech, emotional speech) in order to normalize the tonal space of a speaker (Ladd 2008, 193) into a small number of pitch levels. This approach is exemplified by Pike (1945).

Many prosodic models combine ingredients of both approaches. Autosegmental models, for instance, combine a drastic pitch range normalization (by assuming only two basic pitch levels: low and high) with configurational specifications, describing pitch movements as sequences of pitch targets. In such an approach, pitch contours are characterized as sequences of pitch targets associated with positions defined relative to word stress or prosodic boundaries.

Most intonation models, both configuration- and level-based, refer to global pitch range, although often implicitly. The notion of bottom pitch is a clear example of this. As Ladd (1996, 267) notes, "the *bottom of the speaking range* is a fairly constant feature in an individual's voice." This view is found also in Ladd (2008, 203) and 't Hart (1998, 100). As can be readily observed, listeners are able to detect a speaker's bottom pitch on the basis of phonation characteristics (Honorof and Whalen 2005). As the pitch approaches the bottom of the vocal range, the vibration frequency moves away from the characteristic frequency of the vocal fold vibration; as a result, this vibration may become irregular, as is the case for creaky voice or breathy voice. Creak is not restricted to low-pitched phonation, but its association may be conventional in some language variants or for some speakers.

These considerations regarding pitch range and contour characterization are the basis of the design of a generic transcription system of pitch variation, called Polytonia, which is presented in detail in Mertens (2014). The pitch contour is characterized as a configuration of pitch levels and movements; these levels are defined relative to one another and relative to global pitch range. An intrasyllabic pitch movement is characterized by its starting pitch level, its direction, and its size.

7.4.2 The Proposed Symbolic Tonal Annotation

The tonal transcription specifies pitch level, pitch movement direction, and size for simple or compound movements in a transparent and analytical way. An illustration is given in the second tier of the lower part of figure 7.1, where the symbols L, H, HR, and HF, respectively, indicate a low pitch level, a high pitch level, a large rise starting from a high pitch level, and a large fall starting from a high pitch level. More generally, for each syllable, the transcription indicates the pitch level at vowel onset and, when present, the pitch movement in the rhyme. Complex pitch movements require two or more movement codes, for example, LRF indicates a rise-fall movement starting from a low pitch.

7.4.2.1 Pitch levels The Polytonia annotation defines pitch levels in two ways: locally, as in relative to the context, and globally, as in relative to the speaker's total pitch range. Pitch levels low (L), mid (M), and high (H) are defined locally, relative to one another. A sufficiently large pitch change or movement between a given syllable and some other syllable in the context triggers a change of pitch level. The global interpretation of pitch level, relative to the speaker's pitch range, results in levels top (T) and bottom (B). These additional levels can be explained by their specialized function in speech. In many languages, the bottom pitch marks the end of a maximal prosodic unit or the end of a speaker turn; similarly, the top level is mostly reserved for emotional or emphatic speech portions.

Given this relative successive definition of pitch levels, syllables with the same pitch level code but occurring at different points in the utterance may differ considerably as to their actual F0, provided there are local pitch changes explaining these differences. The overall trend of decreasing pitch, frequently observed in read speech and usually called *declination*, is accounted for in the same way.

7.4.2.2 Pitch intervals Pitch range varies considerably between speakers and speech styles. Hence, pitch interval categories should be expressed relative to pitch range.

The number of pitch interval categories depends on the descriptive model. Many models distinguish between large and small pitch intervals, where the latter typically occur in downstepping or upstepping (IPO model: 't Hart, Collier, and Cohen 1990; INTSINT [International Transcription System for Intonation] model: Hirst and Di Cristo 1998; rhythm and pitch model: Dilley et al. 2006).

The Polytonia transcription system distinguishes two pitch interval sizes: large and small. The size of the small pitch interval is set to 3 ST, which is slightly larger than intrinsic microprosodic variations (Rossi et al. 1981). The large interval is set relative to the observed pitch range (see Mertens 2014, section 5.5, table 2).

7.4.2.3 Intrasyllabic pitch movements There are basically two ways to transcribe a pitch movement within a syllable: as a sequence of pitch levels or as a combination of a pitch level and movement type. For instance, a large rise starting on a low pitch level and ending on a high pitch level will be transcribed respectively as LH or LR, where R represents a large rise. The first approach is common in autosegmental models, which indicate one or more levels (tones) for a syllable and use only two levels (L and H) in combination with downstep and upstep. The limitations of this approach are discussed in Mertens (2014, section 3.2). The second approach is found in some American structuralism works (e.g., Smalley 1964) and is reminiscent of the characterization of contours in the British school (e.g., Crystal 1969; Cruttenden 1986). It allows for a concise and readable representation indicating the pitch level at the start of a syllable, followed by any successive pitch movements in that syllable.

7.4.2.4 Symbol inventory The Polytonia transcription indicates four tonal aspects for each syllable: its starting pitch level, the presence of a simple or compound pitch movement, the shape of each movement (level, rising, or falling), and its size (large or small). All this information can be conveyed by a fairly small set of symbols consisting of five symbols for pitch level (L, H, M, T, and B) and five symbols far large and small pitch movements: R (large rise), F (large fall), r (small rise), f (small fall), and _ (flat). Compound movements use a sequence of the movement symbols, possibly including a level part: for example, RF (rise-fall), _R (level-rise), R_ (rise-level). Although the annotation can represent compound intrasyllabic pitch movements of any complexity, such movements are fairly rare in intonation languages, even in spontaneous speech.

Provided F0 is detected for a given syllable, its pitch movement is always identified, because it can be interpreted on the basis of the stylized pitch in the syllable itself and the pitch range. Pitch level, however, may not be detected for a given syllable. In such cases the pitch movement is shown without pitch level—for instance, R and RF.

For conciseness, a shorthand notation is used. When pitch movement is level and simple, the symbol indicating the level movement is skipped: H_ is simplified to H, M_ is noted M, whereas H_R and HR_ are noted as such (in this example, the two forms distinguish a late rise from an early one). A level movement with missing pitch level, which initially appears as _, is skipped altogether.

When the pitch level starts from L, M, or H but reaches B or T within the same syllable, this is also indicated—for instance LF,B (large low fall to bottom), Lf,B (small low fall to bottom), and HF,B (large high fall to bottom). In many languages, an intrasyllabic fall ending at the bottom level acts as a terminal boundary in the same way as a flat syllable starting at the bottom level. A terminal boundary ends the largest prosodic domain. Most descriptions of prosody observe that declarative utterances end at the bottom pitch level, suggesting that the function of the bottom level may be common to a large number of languages. Similarly, a syllable that rises to the top level produces the same effect as one starting at that level. The algorithm for the automatic transcription of pitch levels and movements is described in detail in Mertens (2014, sections 5.6 and 5.7).

The next section discusses a more advanced prosody transcription, which takes into account stress and prosodic boundaries, makes explicit prosodic structure, and reduces pitch events to a small set of distinctive forms, applying a language-specific intonation grammar.

7.5 Automatic Symbolic Transcription of Intonation

There is more to intonation than pitch variation. Pitch is just one aspect of prosody, next to duration, loudness, rhythm, and so on. In addition, the communicative function of a pitch event depends on where it occurs in the utterance (more precisely, where it occurs in syntactic structure). For instance, depending on its position, a high pitch or a rise may mark focus or the end of a discourse unit. In linguistic analyses of intonation, the relevant positions are generally defined relative to stress and to boundaries.

In articulatory phonetics, a syllable is stressed when it stands out from its context by its increased phonation effort, articulatory strength, by a pitch change, greater loudness, longer duration, or a combination of these features. (For a similar definition, see Ladd 1996, 58.) Difficulties with the definition of stress led to the introduction of related notions, such as accent, pitch accent (Bolinger 1958), and prominence. Most contemporary intonation grammars refer to the notion of stress or to prominence.

Traditionally, a distinction is made between word stress and sentence stress. *Word stress* (also known as lexical stress; Garde 1968) indicates the syllable within a lexical item that will be stressed when the word is pronounced by itself. Out-of-context pronunciation reveals the virtual stress position within the lexical item. Word stress is a language-dependent property: compare English “photography” and French “photographie.” *Sentence stress*, in contrast, indicates a syllable that in an utterance’s pronunciation is actually stressed, hence is marked by some of the cues mentioned herein. In utterances of several words, some syllables carrying word stress will be stressed and others unstressed.

Prosodic structure designates an organization of the speech chain, which identifies structural positions, determined by the presence of stress or prosodic boundaries. These positions are essential for intonation description because the function of a pitch event may be ambiguous when its position is unspecified. Given this underlying structure, the intonation of an utterance is a concatenation of one or more stressed syllables, prosodic boundaries, and unstressed portions between them. Particular sequences form a prosodic unit, such as an “intonation phrase.” Phonological models usually distinguish several layers of hierarchical structuring with their corresponding prosodic units. (Empirical observation from manual and automatic transcription of prosody in colloquial speech suggests that some of these layers are highly speculative.)

Given the fact that prosodic structure refers to stress and indirectly to word stress, which is language-dependent, an intonation model will inevitably be specific to a given language and so will be the corresponding representation and the automatic procedure to compute it.

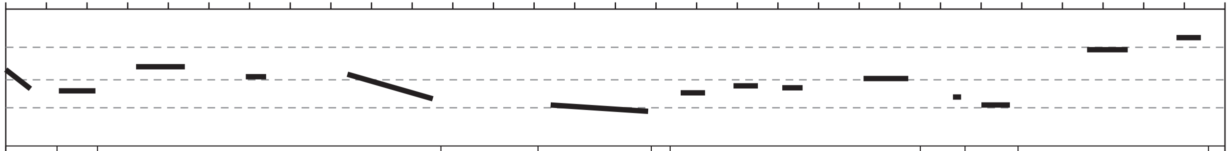
Having clarified the notions of prosodic structure, stress, and intonation units, we now turn to automatic decoding of this structure. To obtain a comprehensive transcription of intonation, specifying structural positions in the utterance as well as the pitch events occurring for them, information is needed about the location of stressed syllables and prosodic unit boundaries.

In what follows, I illustrate the automatic computation of such a transcription for French, first however summarizing French prosodic structure.

Two layers of organization may be identified in French intonation (Mertens 2006; Mertens et al. 2001). On a first layer, the stress group consists of a lexical item carrying word stress and the adjacent words (mostly grammatical words) that syntactically depend on it. So stress groups may be derived from word stress and syntactic dependency relations between words. Each stress group contains a syllable with virtual stress. On a hierarchically higher layer, the intonation group consists of a sequence of one or more stress groups, in which the last carries sentence stress, whereas the other virtual stress positions remain unstressed. For the phonological characterization of distinctive pitch contours observed on intonation groups in French, three structural positions are needed: the final stressed syllable, the penultimate, and the syllable carrying initial stress. The latter position is optional within the intonation group. When present, it usually coincides with the initial syllable of a lexical item, rarely with the second, and never with final stress.

Speech corpora that provide (temporally aligned) annotations for word forms, part of speech, sounds, syllables, stress (or prominence), and a pitch labeling (such as Polytonia) allow for the computation of pitch contours, specifying their structural positions and the pitch events (pitch level, pitch movement, or tone) associated with these positions. The resulting annotation is illustrated in figure 7.6, which shows the input annotation tiers just mentioned and the output: the structural positions in the pitch contour with their respective labels. The same information could be represented by association lines connecting tones to their position in the segmental layer, in combination with stress and boundary marks, as is common in autosegmental approaches. The structural positions are indicated as follows. Stress is marked by the IPA stress mark preceding the pitch level; for instance ˈH indicates a syllable with final stress, on a high pitch level. Initial stress is differentiated from final stress by the suffix *i* appended to the pitch symbol—for instance ˈMi. The penultimate position is noted by the suffix *p*, as in Lp. Finally, X is used when some input information is missing (such as the pitch level on the penultimate of *simplifie*) or for syllables with a special status (such as the schwa following the stressed syllable in *courbe*).

The transcription of a longer fragment appears in figure 7.7, with just two annotation layers: word form and intonation transcription.



donc	on	simplifie			_	euh	l’	évolution			de	la	courbe		_
dɔ̃	kɔ̃	sɛ̃	pli	fi	_	ø	le	vɔ	ly	sjɔ̃	də	la	kuʀ	bə	_
		p		P						p			P		
HF	L	M		Mf		L	L	L	L	H		L	H	H	
		’Mi	X	’M					Lp	’H		Lp	’H	X	

Figure 7.6
Input and output for the automatic transcription of intonation contours for a short sample of the C-Prom corpus (Avanzi et al. 2010) of French, file cnf-be, which records an academic presentation by a female speaker. The utterance is “Donc on simplifie euh l’évolution de la courbe” (So we simplify the shape of the curve). The annotation tiers indicate (i) word form, (ii) syllable, (iii) prominence (*P* and *p* indicate strong and weak prominence, respectively), (iv) pitch level and movement (obtained by Polytonia), (v) the intonation groups, and (vi) structural positions within the pitch contour with their respective pitch labels. In tier 6, final stress is marked by an IPA stress mark preceding the pitch symbol, for example, ˈH; initial stress is also indicated by a stress diacritic and an *i* suffix appended to the pitch symbol, for example, ˈMi; the penultimate syllable is noted by the suffix *p*, for example, Lp. The *X* symbol indicates missing information or a syllable with a special status.

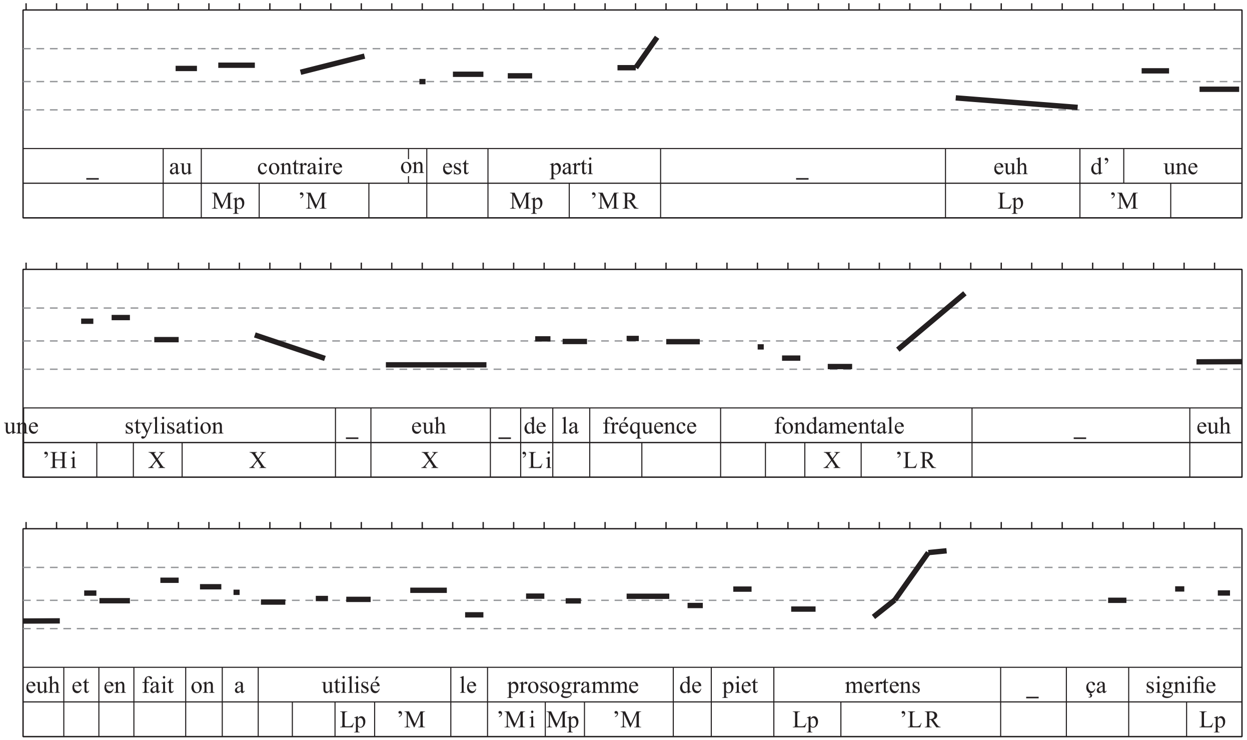


Figure 7.7
Automatic transcription of intonation units and contours for a fragment of the C-Prom corpus, recording "cnf-be," pronounced by a female speaker. The annotation tiers indicate word form and intonation contour transcription using the conventions of figure 7.6. The passage translates to: "On the contrary, we started from a stylization of fundamental frequency. And in fact we used the Prosogram of Piet Mertens. This means . . ."

The system for automatic transcription of intonation in French, illustrated in figures 7.6 and 7.7, was elaborated in collaboration with Sandrine Brognaux (2015) and is called ToPPos-Fr, for "tones and prosodic positions in French." It uses several inputs (information sources), which may be either specified explicitly by annotation tiers or computed by dedicated modules on the basis of additional information sources. For instance, prominence information may be provided by the corpus annotation (as is the case for the C-Prom corpus used in the figures) or by a prominence detection module using acoustic cues (Brognaux 2015). Similarly, stress groups may be determined using part-of-speech information, together with local syntactic relations (or, alternatively, from the enumeration of part-of-speech sequences that form stress groups). Two further steps follow: one for the identification of structural positions and the other for mapping pitch levels and movements onto the set of tonal categories that the intonation model uses. The next two paragraphs provide more details on these steps.

For identifying the three structural positions of final, penultimate, and initial stress, a set of rules is applied to each syllable. Final stress position is assigned to a prominent syllable carrying word stress, that is, when it is the final syllable of a stress group and that syllable does not contain a schwa vowel. Final stress position then forms the right boundary of the corresponding intonation group. An optional schwa may occur after final stress within the same word (as is the case for the word *courbe* in the example in figure 7.6); this schwa then forms the last syllable of the intonation group but is not itself in final stress position. Initial stress position is inferred when a prominent syllable is word-initial but not in final stress position. The syllable preceding a final stress is labeled "penultimate." Additional rules deal with special cases. For instance, the last syllable of a stress group that is followed by a pause (or by a schwa and a pause) and reaches the bottom pitch level is considered to be in final stress position, whether it is prominent or not.

As a final step, pitch levels and movements (as identified by Polytonia) are mapped onto the categories used in the intonation model targeted by the transcription system. Tier 2 of figure 7.7 illustrates one such model or transcription convention: ToPPos-Fr. More generally, a mapping may entail a reduction of the number of retained pitch levels, a simplification of compound movements to simple ones, or a conversion of pitch movements to sequences of pitch levels. Moving categorization even further, the fairly large set of observed pitch events may be mapped onto a small set of functional categories such as major continuation, minor continuation, terminal contour, question contour, and focus, and so forth. Whatever transcription convention is selected, the system architecture remains unchanged; only the mapping rules differ.

The system for automatic transcription of intonation contours in French outlined in this section simulates perceptual and cognitive processing. When we decode the intonation of an utterance, we combine multiple sources of information, both perceptual and grammatical (our knowledge of the language grammar). This interaction process is simulated by a rule-based module.

7.6 Conclusion

Linguistic analysis of intonation aims at identifying the distinctive prosodic forms of spoken language. It studies their alignment with segmental structures, be it syllables, lexical items, syntactic structures, or discourse units. It also investigates the meaning of the primitive forms, both taken separately and in interaction with other linguistic structures (syntactic constructions, discourse structure). This chapter is limited to the formal aspects of prosody: the auditory forms used by intonation.

The central assumptions of our approach may be summarized as follows.

Perceptual processing in the auditory system affects the way listeners interpret F0 variations in speech. Spectral instability introduces a segmentation into the speech signal and, hence, in the evolution of its acoustic prosodic parameters. The resulting representation in auditory memory consists of a chain of elements that correspond to local maxima of sonority with relative spectral stability and correspond roughly to the syllabic nuclei.

The proposed type of pitch stylization simulates the impact of this segmentation on the perception of pitch in speech. In the stylized contour, a syllable's pitch is either static (flat) or dynamic (when it includes an audible pitch glide), and the evolution of pitch during speech is characterized by the melodic intervals that separate successive syllables or occur within individual syllables. This approach aspires to modeling psychological reality and cognition while at the same time providing a model of acoustic F0 contours that has been applied successfully in resynthesis, text-to-speech synthesis, and intonation transcription.

In linguistic decoding of speech prosody by humans, pitch variations are normalized relative to the global pitch range of the individual speaker. The size of pitch steps and movements is evaluated relative to the speaker's pitch span. The bottom and top of his pitch range are partially identified from phonation properties.

In many transcriptions and models of prosody used in linguistics, the evolution of pitch is represented by pitch levels and pitch movements. These may be computed automatically, starting from the pitch stylization based on tonal perception, as is illustrated by the Polytonia program (section 7.4).

Comprehensive linguistic models of intonation postulate a prosodic structure, which takes into account stress and boundaries, defines prosodic units (such as the intonation phrase), and characterizes pitch contours as sequences of pitch events (tones) associated with positions in prosodic structure. The actual grammar will differ from one language to the next. By combining multiple information sources, such intonation contour specifications may be derived automatically, as illustrated by the program set out in section 7.5.

The latter application provides a specification of intonation matching the requirements of a so-called *phonological intonation model*. It posits a set of abstract, symbolic primitives (pitch levels and movements) that combine with positions (stress, prestress, emphatic stress, boundaries, and so on) in prosodic structure to generate well-formed intonation contours, as well as their acceptable phonetic instantiations (by synchronizing target points with syllable structure).

Pitch levels and movements may be viewed as the phonemes of prosody, defined relative to one another and without meaning of their own. An intonation morpheme is formed when one or more phonemes are associated with a given position in prosodic structure, such as stress. Examples of such morphemes in French are (using the notation of section 7.5) ˈH, ˈM, ˈHF, ˈLR, ˈHi, ˈB, and Hp. Such intonation morphemes do have a meaning and may form minimal pairs. The sequence of morphemes within a prosodic unit forms an intonation contour, which ultimately may be viewed as a syntactic constituent of the prosodic domain. In this view, the concepts of phoneme, morpheme, and phrase in the prosodic domain parallel those of the segmental domain.

Meaning or function of intonation contours is not dealt with in this chapter, which focuses on bottom-up analysis, recognition, and transcription of the prosodic primitives and their organization into larger units. Propositions for a compositional characterization of intonation meaning in French and for the interface with syntactic constructions are in Mertens (2006). As far as intonation meaning is concerned, formal, objective, reproducible methods appear to be scarce.

The output representations (stylization, symbolic pitch levels and movements, intonation transcription) derived by the model have been used as input to a text-to-speech system (Mertens et al. 2001). In this respect, our approach is related to the analysis by synthesis approach that Hirst (2011) proposed.

In his critical review of prosody research methodology, Xu (2011) distinguishes several approaches: analysis by transcription, analysis by introspection, analysis by hypothesis testing, and analysis by modeling. Our approach belongs mainly to the last type. Xu (2011, section 5) writes: "Potentially the most rigorous test of our understanding of prosody, especially in terms of predictive knowledge, is computational modeling. . . . Its importance has in general not been duly recognized." Because it combines acoustic, perceptual, and linguistic representations of prosody, our approach may also be characterized as an integrated model of prosody. It provides a multidisciplinary synthesis of methods and insights from auditory (manual) transcription of intonation, articulatory phonetics, acoustic phonetics, psychoacoustics, phonology, linguistics, speech processing, computing, and natural language processing.

Independent of such methodological considerations, Prosogram provides a tool that enables researchers to easily obtain a quantified, objective, and perceptually motivated representation of prosody in continuous speech (Mertens 2020). It is our hope that the availability of this tool will encourage researchers not to take prosodic theories for granted, but to discover the richness of prosody, as it emerges from large-scale corpora of spontaneous speech.

References

Avanzi, M., A. C. Simon, J.-Ph. Goldman, and A. Auchlin. 2010. "C-PROM: An Annotated Corpus for French Prominence Studies." In *Proceedings of Prosodic Prominence: Perceptual and Automatic Identification, Speech Prosody 2010 Workshop*. https://www.isca-speech.org/archive/sp2010/papers/sp10_2005.pdf.

Baken, R. J., and R. F. Orlikoff. 2000. *Clinical Measurement of Speech and Voice*. San Diego: Singular Publishing.

Bardiaux, A., and P. Mertens, 2014 "Normalisation des contours intonatifs et étude de la variation régionale en français." *Nouveaux cahiers de linguistique française* 31:273–284.

Barnes, J., A. Brugos, N. Veilleux, and S. Shattuck-Hufnagel. 2014. "Segmental Influences on the Perception of Pitch Accent Scaling in English." In *Proceedings of Speech Prosody 7*, edited by N. Campbell, D. Gibbon, and D. Hirst, 1125–1129. https://www.isca-speech.org/archive/SpeechProsody_2014/pdfs/219.pdf.

Boersma, P., and D. Weenink. 2012. *Praat: Doing Phonetics by Computer*. Version 5.4.18. Computer program. http://www.praat.org/.

Bolinger, D. L. 1951. "Intonation: Levels vs. Configurations." *Word* 7:199–210.

Bolinger, D. L. 1958. "A Theory of Pitch Accent in English." *Word* 14:109–149.

Brognaux, S. 2015. "Expressive Speech Synthesis: Research and System Design with Hidden Markov Models." PhD diss., Université catholique de Louvain and Université de Mons.

Christodoulides, G., and M. Avanzi. 2014. "An Evaluation of Machine Learning Methods for Prominence Detection in French." In *Proceedings of Interspeech*. https://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_0116.pdf.

Christodoulides, G., and A. C. Simon. 2015. "Exploring Acoustic and Syntactic Cues to Prosodic Boundaries in French: A Multi-Genre Corpus Study." In *Proceedings of the International Congress of Phonetic Sciences*. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/proceedings.htlm.

Cruttenden, A. 1986. *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

d'Alessandro, C., and M. Castellengo. 1994. "The Pitch of Short-Duration Vibrato Tones." *Journal of the Acoustical Society of America* 95:1617–1630.

d'Alessandro, C., and P. Mertens. 1995. "Automatic Pitch Contour Stylization Using a Model of Tonal Perception." *Computer Speech and Language* 9 (3): 257–288.

De Looze, C., and D. J. Hirst. 2010. "Integrating Changes of Register into Automatic Intonation Analysis." In *Proceedings of the Speech Prosody Conference*. https://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_0116.pdf.

De Looze, C., and D. Hirst. 2014. "The OMe (Octave-Median) Scale: A Natural Scale for Speech Melody." In *Proceedings of the Speech Prosody Conference*.

de Moraes, J. A., and A. Rilliard. 2014. "Illocution, Attitudes and Prosody." In *Spoken Corpora and Linguistic Studies*, edited by T. Raso and H. Mello, 233–270. Amsterdam: John Benjamin.

Dilley, L., M. Breen, E. Gibson, M. Bolivar, and J. Kraemer. 2006. "A Comparison of Inter-Coder Reliability for Two Systems of Prosodic Transcriptions: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices)." In *Proceedings of the International Conference on Spoken Language Processing*. http://speechlab.cas.msu.edu/PDF/Old Conference Proceedings/Dilley Breen Bolivar Kraemer Gibson 2006.pdf.

Garde, P. 1968. *L'accent*. Paris: Presses Universitaires de France.

Goldman, J.-Ph., A. Auchlin, A. C. Simon, and M. Avanzi. 2007. "Phonostylographe: Un outil de description prosodique. Comparaison du style radiophonique et lu." *Nouveaux Cahiers de Linguistique Française* 28:223–241.

Goldman, J.-P., M. Avanzi, A. Lacheret-Dujour, A. Simon, and A.-C. Auchlin. 2007. "A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French." In *Proceedings of Interspeech*. https://www.isca-speech.org/archive/interspeech_2007/i07_0098.html.

Grice, M. 2006. "Intonation." In *Encyclopedia of Language and Linguistics*, 2nd ed., edited by K. Brown, vol. 5, 778–788. Oxford: Elsevier.

Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Heldner, M., and M. Włodarczak. 2015. "Pitch Slope and End Point as Turn-Taking Cues in Swedish." In *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0307.pdf.

Hermes, D. J. 1987. "Vowel-Onset Detection." *IPO-Annual Progress Report* 22:15–24.

Hermes, D. J. 2006. "Stylization of Pitch Contours." In *Methods in Empirical Prosody Research*, edited by S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schließer, 29–61. Berlin: Walter de Gruyter.

Hermes, D. J., and J. C. Van Gestel. 1991. "The Frequency Scale of Speech Intonation." *Journal of the Acoustical Society of America* 90:97–102.

Hirst, D. J. 2011. "The Analysis by Synthesis of Speech Melody: From Data to Models." *Journal of Speech Science* 1 (1): 55–83.

Hirst, D. J., and A. Di Cristo. 1998. "A Survey of Intonation Systems." In *Intonation Systems: A Survey of Twenty Languages*, edited by D. Hirst and A. Di Cristo, 1–44. Cambridge: Cambridge University Press.

Hirst, D. J., A. Di Cristo, and R. Espesser. 2000. "Levels of Representation and Levels of Analysis for Intonation." In *Prosody: Theory and Experiment*, edited by M. Horne, 51–87. Dordrecht: Kluwer Academic Publishers.

Hirst, D. J., P. Nicolas, and R. Espesser. 1991. "Coding the F0 of a Continuous Text in French: An Experimental Approach." In *Proceedings of the International Congress of Phonetic Sciences*, 234–237. Aix-en-Provence: Université de Provence. Service des publications.

Honorof, D. N., and D. H. Whalen. 2005. "Perception of Pitch Location within a Speaker's F0 Range." *Journal of the Acoustical Society of America* 117 (41): 2193–2200.

House, D. 1990. *Tonal Perception in Speech*. Lund, Sweden: Lund University Press.

House, D. 1995. "The Influence of Silence on Perceiving the Preceding Tonal Contour." *Proceedings of the International Congress on Phonetic Sciences 13* vol. 1, 122–125. https://www.internationalphoneticassociation.org/icphs/icphs1995.

House, D. 1996. "Differential Perception of Tonal Contours through the Syllable." In *Proceedings of the International Conference of Spoken Language Processing*, 2048–2051. http://www.asel.udel.edu/icslp/cdrom/vol4/064/a064.pdf.

Jassem, W., and K. Kudela-Dobrogowska. 1980. "Speaker-Independent Intonation Curves." In *The Melody of Language: Intonation and Prosody*, edited by L. R. Waugh and C. H. van Schooneveld, 135–148. Baltimore: University Park Press.

Lacheret-Dujour, A., S. Kahane, and P. Pietrandrea, eds. 2019. *Rhapsodie. A Prosodic and Syntactic Treebank for Spoken French*. Amsterdam: Benjamins.

Ladd, D. R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. R. 2008. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Mertens, P. 2004a. "Un outil pour la transcription de la prosodie dans les corpus oraux." *Traitement automatique des langues* 45 (2): 109–130.

Mertens, P. 2004b. "The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model." In *Proceedings of Speech Prosody 2004*, edited by B. Bel and I. Marlien. Aix-en-Provence: Laboratoire Parole et Langage.

Mertens, P. 2006. "A Predictive Approach to the Analysis of Intonation in Discourse in French." In *Prosody and Syntax*, edited by Y. Kawaguchi, I. Fonagy, and T. Moriguchi, 64–101. Amsterdam: John Benjamins.

Mertens, P. 2013. "Automatic Labelling of Pitch Levels and Pitch Movements in Speech Corpora." In *Proceedings TRASP 2013, Tools and Resources for the Analysis of Speech Prosody*, edited by B. Bigi and D. Hirst, 42–46. http://www2.lpl-aix.fr/~trasp/Proceedings/TRASP2013_proceedings.pdf.

Mertens, P. 2014. "Polytonia: A System for the Automatic Transcription of Tonal Aspects in Speech Corpora." *Journal of Speech Sciences* 4 (2): 17–57.

Mertens, P. 2019. "From Pitch Stylization to Automatic Tonal Annotation of Speech Corpora." In *Rhapsodie. A Prosodic and Syntactic Treebank for Spoken French*, edited by A. Lacheret-Dujour, S. Kahane, and P. Pietrandrea, 233–250. Amsterdam: Benjamins.

Mertens, P. 2020. *Prosogram User's Guide*. https://sites.google.com/site/prosogram.

Mertens, P., F. Beaugendre, and Ch. d'Alessandro. 1997. "Comparing Approaches to Pitch Contour Stylization for Speech Synthesis." In *Progress in Speech Synthesis*, edited by J. P. H. vanSanten, R. W. Sproat, J. P. Olive, and J. Hirschberg, 347–363. New York: Springer-Verlag.

Mertens, P., J.-Ph. Goldman, É. Wehrli, and A. Gaudinat. 2001. "La synthèse de l'intonation à partir de structures syntaxiques riches." *Traitement Automatique des Langues* 42 (1): 145–192.

Mertens, P., and A. C. Simon. 2013. "Towards Automatic Detection of Prosodic Boundaries in Spoken French." In *Proceedings of the Prosody-Discourse Interface Conference*, edited by P. Mertens and A. C. Simon, 81–87. https://www.arts.kuleuven.be/ling/cohistal/conference/idp2013/documents/proceedings_idp2013.

Nolan, F. 2003 "Intonational Equivalence: An Experimental Evaluation of Pitch Scales." In *Proceedings of the International Congress of Phonetic Sciences*, 771–774. https://www.internationalphoneticassociation.org/icphs/icphs2003.

Patel, A. D., J. R. Iversen, and J. C. Rosenberg. 2006. "Comparing the Rhythm and Melody of Speech and Music: The Case of British English and French." *Journal of the Acoustical Society of America* 119:3034–3047.

Pike, K. L. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Poiré, Fr., and S. Kaminskaia. 2004. "Comparing Intonation of Two Varieties of French Using Normalized F0 Values." In *Proceedings of the Eighth International Conference on Spoken Language Processing*, 1305–1308. https://www.isca-speech.org/archive/archive_papers/interspeech_2004/i04_1305.pdf.

Rietveld, T., and P. Vermillion. 2003. "Cues for Perceived Pitch Register." *Phonetica* 60 (4): 261–272.

Rossi, M. 1971. "Le seuil de glissando ou seuil de perception des variations tonales pour la parole." *Phonetica* 23:1–33.

Rossi, M. 1978. "Interactions of Intensity Glides and Frequency Glissandos." *Language and Speech* 21:384–396.

Rossi, M., A. Di Cristo, D. Hirst, Ph. Martin, and Y. Nishinuma. 1981. *L'intonation. De l'acoustique à la sémantique*. Paris: Klincksieck.

Scheffers, M. T. M. 1988. "Automatic Stylization of F0-Contours." In *Proceedings of the Seventh FASE Symposium*, edited by W. A. Ainsworth and J. N. Holmes, 981–987. Edinburgh: Institute of Acoustics.

Smalley, W. A. 1964. *Manual of Articulatory Phonetics*. New York: Practical Anthropology.

Spaai, G. W. G., A. Storm, A. S. Derksen, D. J. Hermes, and E. F. Gigi. 1993. *An Intonation Meter for Teaching Intonation to Profoundly Deaf Persons*. IPO Manuscript 968. Eindhoven: Technische Universiteit Eindhoven.

Taylor, P. 1994. “The Rise/Fall/Connection Model of Intonation.” *Speech Communication* 15 (1/2): 169–186.

Taylor, P. 2000. “Analysis and Synthesis of Intonation Using the Tilt Model.” *Journal of the Acoustical Society of America* 107 (3): 1697–1714.

’t Hart, J. 1998. “Intonation in Dutch.” In *Intonation Systems: A Survey of Twenty Languages*, edited by D. Hirst and A. Di Cristo, 96–111. Cambridge: Cambridge University Press.

’t Hart, J., R. Collier, and A. Cohen. 1990. *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press.

Xu, Y. 2005. “Speech Melody as Articulatorily Implemented Communicative Functions.” *Speech Communication* 46:220–251.

Xu, Y. 2011. “Speech Prosody: A Methodological Review.” *Journal of Speech Sciences* 1 (1): 85–115.

8

The Kiel Intonation Model—KIM

Oliver Niebuhr

8.1 Introduction

The Kiel intonation model (KIM) was developed as a contour model for Northern Standard German in the late 1980s by Klaus Kohler and his colleagues at the former Kiel Institute of Phonetics and Digital Speech Processing. It has since been refined and extended in a number of ways. These refinements and extensions rely on a growing body of empirical evidence from perception experiments and supplementary acoustic analysis of dialogue recordings, mainly conducted since 2003 by Benno Peters, Tamara Rathcke, Ernst Dombrowski, Klaus Kohler, Rabea Landgraf, and the author of the present chapter.

The ultimate goal of the KIM is to determine all communicative functions of Standard German prosody and relate them to distinctive prosodic features, derived from detailed descriptions, and their contextual as well as individual variation. The wording "all communicative functions" implies that the KIM does not differentiate between so-called linguistic and paralinguistic meanings in order to give the former priority over the latter. Rather, all types of meanings or functions from information structure to attitudinal meanings and argumentation structure to emphatic intensification are taken into account on an equal footing (Kohler 2007).

At the heart of the KIM is a phonological system. It is probably the only system to date that explicitly includes phonological levels of perceptual prominence, next to the phonological categories of intonation and phrasing that are offered by many other models as well. But the KIM as a whole is even larger. It also takes phenomena such as speech rate, pitch register, and types of disfluencies into account that often extend across several prosodic phrases. However, this chapter focuses on phonological categories that are associated with the domain of the prosodic phrase; it does so to comply with the maximum number of words and because almost all research and labeling within the KIM framework so far has been concerned with these phrase-level categories.

As this chapter provides the first extensive summary of the KIM after almost twenty years, the revised phonological system differs quite a bit from Kohler's (1991a, 1991b, 1997) early system presentations. Any changes and suggestions made in this context represent the author's own reflection about what the KIM currently looks like and how it should be developed in the future. The current phrase-level phonology of the KIM distinguishes among three prominence levels, nineteen intonation categories, nine types of phrase boundaries, and fourteen strategies for emphatic intensification. Not least in view of these relatively many categories, the target audience of the KIM comprises those who approach prosody from a general communicative perspective, and who would like their annotation to be as rich as possible in terms of both forms and functions.

Attached to the KIM is an inventory for prosodic labeling (PROLAB) that translates the phonological categories of intonation, phrasing, prominence, and emphasis into sequences of simple ASCII letters. These symbols are annotated on a single tier with a conventionalized syntactic structure. However, there are tools that can automatically convert these annotations into multiple-tier representations. These tools ensure that PROLAB annotations are largely compatible and interchangeable across different software and transcription systems such as Praat and Emu (John 2012), which were not as popular when PROLAB was developed.

Finally, the KIM is obviously most suitable for German data; however, a number of phonological distinctions have recently also been found in a range of other languages, for example, English (Kleber 2006) Swedish (Ambrazaitis 2005, 2009), and Estonian (Asu 2006), making the KIM potentially also beneficial for a wider international audience.

8.2 A Brief Retrospect

The origins of the KIM lie in a practical rather than theoretical challenge: to develop a set of categories and rules that are able to generate proper prosodic patterns for German text-to-speech (TTS) synthesis within the Rulesys/Infovox framework (Carlson, Granström, and Hunnicut 1990; Kohler 1997). Accordingly, the first key publication of the KIM by Kohler (1991a, 1991b), which is still the major reference to the model, is structured similar to a generative grammar and develops decision trees with binary features of prominence and intonation (see figure 8.1). These feature trees are complemented by twenty-six pages specifying not fewer than 130 rules on the concatenation and context-sensitive implementation of these binary features.

The focus of the KIM changed with the start of the large international Verbmobil project in the mid-1990s (Wahlster and Engelkamp 1992). Verbmobil was supposed to create an application that both detected anger and disfluencies and also provided its users with instant translations of German, English, and Japanese spontaneous speech, for example, during telephone conversations. Implementing the idea of Verbmobil required recording, annotating, and analyzing what at the time were huge corpora of spontaneous speech. The Kiel Phonetics Institute was part of the Verbmobil network, and so the Kiel Corpus of Spontaneous Speech was recorded. Based on an appointment-making scenario, it is still one of the largest German corpora with full segmental and prosodic annotations at multiple phonetic-phonological levels (Peters 2005).

The KIM was used in the Verbmobil project as a resource for phonology-based prosodic labeling (Kohler 1996). The labeling tool that was developed, PROLAB, took the prominence and intonation categories of the KIM and translated them into machine-friendly seven-bit ASCII labels (Kohler 1997). The work with PROLAB in Verbmobil had a crucial side effect: applying the KIM to hours of spontaneous speech revealed quite a few discrepancies between the categories of the model (i.e., their phonetic definitions or assumed communicative functions) and actual speech data. These inadequacies and insufficiencies became the driving force for refining and extending the model's prosodic phonology (Kohler 2007).

This does not mean, however, that KIM categories were simply changed or added with reference to single-case examples or distributional differences between prosodic patterns in the Kiel corpus. Production data always played second fiddle in the KIM's

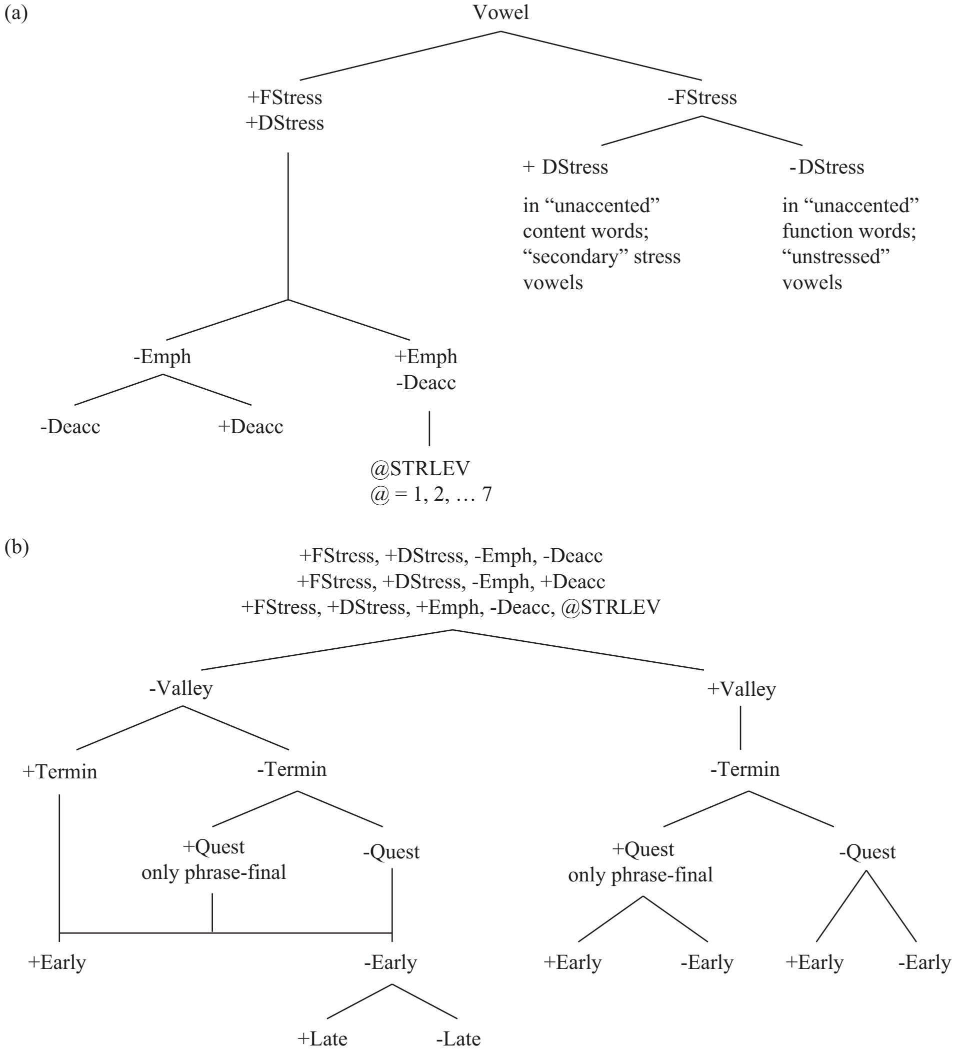

(a)
Vowel
+FStress
+DStress
-FStress
+ DStress
in "unaccented" content words; "secondary" stress vowels
-DStress
in "unaccented" function words; "unstressed" vowels
-Emph
+Emph
-Deacc
-Deacc
+Deacc
@STRLEV
@ = 1, 2, … 7
(b)
+FStress, +DStress, -Emph, -Deacc
+FStress, +DStress, -Emph, +Deacc
+FStress, +DStress, +Emph, -Deacc, @STRLEV
-Valley
+Valley
+Termin
-Termin
-Termin
+Quest
only phrase-final
-Quest
+Quest
only phrase-final
-Quest
+Early
-Early
+Early
-Early
+Early
-Early
+Late
-Late

Figure 8.1

Binary distinctive features of (a) prominence and (b) intonation and their corresponding hierarchical organization in the original version of the KIM. Deacc, deaccented; DStress, duration-dominated stress; Emph, emphasis; FStress, f0-dominated stress; Quest, question; STRLEV, stress level; Termin, terminal F0 fall/rise. *Source*: After Kohler (1991a, 1991b, 1997).

prosodic phonology. The methodological fundamentals of the KIM are rooted in the Prague school (Trubetzkoy 1939), which, in turn, is based on the semiotic theory of de Saussure (1916). Unlike in the distribution-oriented, behavioristic American structuralism (Bloch 1948), a basic principle of de Saussure's theory is that phonological categories have to be set up in contrast to one another *and* with direct evidence that this contrast is associated with a change in meaning. As a result, production data were primarily used in the KIM for generating hypotheses about prosodic form-function relationships. Then it was the critical task of perception experiments to test these hypotheses and, on this basis, to revise or add phonological categories of the prosodic model. The listeners' task in these perception experiments was always to judge a (resynthesized) prosodic parameter space by making direct or indirect reference to meaning (see section 8.3.1).

In line with having its roots in the European tradition of linguistics, the KIM's intonation categories are inspired by the intonation analyses of major British school scholars, for example, Halliday (1967), Jones (1969), and O'Connor and Arnold (1970). Kohler projected their tune structures and some of their paradigmatic oppositions onto German, but reconceptualized them as arising from two basic F0 contour types, each of which can be in different synchronization with the vowel of the accented syllable. The idea about the relevance of synchronization was obviously influenced by the seminal work of Bruce (1977). This becomes even clearer when the following is taken into account. While the KIM is, on the one hand, based on holistic contours, the synchronization-dependent phonological distinctions within each contour type rely, on the other hand, solely on local landmarks within these contours—most importantly F0 minima and maxima (see section 8.4.3).

In summary, having an idea about the KIM's history and development helps in understanding why the KIM is as it is today. Many characteristics that set the KIM apart from other prosodic models were not grounded in a new theoretical approach but rather arose from the fact that the model "grew up" in technical speech application environments, first Rulesys/Infovox and later Verbmobil. These environments motivated the two phonological subsystems, one for prominence and the other for intonation, and determined the PROLAB representation of KIM categories in simple ASCII letters. Furthermore, the intensive engagement with analyzing and processing spontaneous speech over more than twenty years, in combination with adopting the phonological concept of the Prague school, led to a rich, meaning-oriented phonological structure, particularly in terms of intonation.

It can also be considered a consequence of the KIM's technical background that phonological categories have an intermediate degree of abstraction. That is, categories are explicitly defined independently of micro-prosodic and individual variation (Kohler 1991a, 1991b). However, at the same time, specific characteristics of time and frequency serve as distinctive features and are an essential part of the categories' definitions. These features were obviously inspired by both traditional contour models and the tonal-target approach (in the sense of the early work of Bruce). Accordingly, the KIM distinguishes basic contour types, but sets up categories within each contour type on the basis of local F0 landmarks. By this hybrid approach to intonation, the KIM clearly differs from pure contour models such as the Dutch IPO model, named after the Eindhoven Instituut voor Perceptie Onderzoek (Eindhoven Institute for Perception Research). In this author's opinion, the KIM could be called a "pattern model."

8.3 Establishing Phonological Structures

8.3.1 The KIM's Methodological Guideline

Insights and progress in the field of segmental phonology essentially rely on the fact that we know where the boundaries between meaning-related units are and what their communicative (semantic or grammatical) function is. We can make use of this knowledge as a point of reference when we set up a system of paradigmatic contrasts and study variation within and between the categories of this system, for example, in the form of coarticulation and reduction processes.

The starting point in prosodic phonology, and of intonation in particular, is a completely different one (see the discussions in Hirschberg 2002, 2004). The general (and still widely unresolved) uncertainty about the boundaries between intonational elements is most prominently reflected in the controversy about contour and tonal-target (i.e., level) models (Ladd 1996). Moreover, the difference in meaning conveyed by an audible difference in intonation is often rather difficult to pinpoint, for two reasons. First, intonational meanings are typically quite abstract, and second, these meanings can take different shades in different conversational contexts. It may also be that the second fact is a consequence of the first.

The reason that intonational meanings are "essentially pragmatic in nature" (Hirschberg 2004, 515) could also be related to the kind of meaning conveyed. Intonational meanings are often attributive in the sense that they ascribe a certain evaluative quality to a piece of information, the dialogue partner, or the situation in general. A similar kind of attributive meaning is conveyed by lexical adjectives, and just like intonational meanings, the meanings of lexical adjectives such as *funny, crazy, odd* are also far from straightforward and independent of contextual interpretation. For instance, the synonyms for *funny* given in the *Oxford English Dictionary* (Winchester 2003) range from *humorous* through *unpleasant* and *wrong* to *strange* and *curious*.

Many intonational meanings that were claimed to be nonattributive at first sight turned out to be attributive on closer inspection. A prime example is the difference between sentence-final rising and falling intonation, for instance, in German and Swedish (Petrone and Niebuhr 2013; Strömbergsson, Edlund, and House 2012). This difference was once firmly associated with signaling sentence mode. Today, we know that the rise-versus-fall difference actually conveys an attitudinal meaning difference, indicating whether the speaker super- or subordinates himself or herself to the interlocutor (Kohler 2004). This new meaning difference was suddenly able to account for a lot of mismatches between intonation and sentence mode that had puzzled researchers for quite a long time.

So, intonational meaning variation may appear a bit less exotic if we compare it to the range of meanings expressed by lexical adjectives. Moreover, approaching intonational meanings from an attributive perspective, which goes beyond the "usual suspects" of sentence mode, information, and discourse structure, can make the definition of intonational meanings a more worthwhile endeavor, even if not all intonational meanings turn out to be attributive in the end. Against this background, it is a key assumption of the KIM that phonological categories have meanings, that these meanings can be identified—or at least roughly outlined in contrast to one another—and that on this basis, meanings can be points of reference in establishing a prosodic phonology, analogous to the segmental phonology of the Prague school.

Accordingly, intonational meanings were always determined in the KIM framework in parallel to establishing intonational contrasts. The meaning difference of an intonational contrast had already been the subject of multiple studies, before the specific prosodic characteristics of the categories involved were investigated. In other words, the phonological structures of the KIM were developed according to the principle that "form follows function, not function form" (Xu 2004, 91).

The form-follows-function principle, in the chronological sense of understanding a function before understanding how it is encoded in the speech signal,was implemented on the basis of perception experiments that had the same underlying concept in all cases. First, a prosodic parameter space (i.e., a stimulus continuum) was generated on the basis of a short utterance, typically by means of pitch synchronous overlap and add (PSOLA) resynthesis in Praat (Boersma 2001) or, until the late 1990s, based on the Rulesys TTS system for which the KIM was originally developed (Kohler 1987, 2007).

Varied parameters were, for example, F0-peak synchronization or scaling, the duration of the accented syllable, or the direction and steepness of the phrase-final F0 movement. Then these stimuli were presented in multiple randomized repetitions to groups of naive German listeners. The listeners made direct or indirect judgments about the intonational/prosodic meanings in the stimuli.

8.3.2 Perception Experiments Based on Direct Judgments about Intonational Meanings

Direct judgments are obtained by means of the semantic-differential paradigm (Osgood, Suci, and Tannenbaum 1957). That is, during its occurrences in the experiment, each stimulus is judged on five to nine bipolar scales whose opposite adjectives at the scale ends are selected to cover three semantic dimensions: arousal, potency, and valence. Most studies, such as those of Dombrowski (2003), Dombrowski and Niebuhr (2010), Kohler (2005, 2011), and Niebuhr et al. (2012), used a seven-point Likert scale. However, a number of recent studies (e.g., Niebuhr 2008) also used continuous scales (visual analogue scales), Rietveld and Chen (2006) having shown that these scales trigger more distinct judgments and are more sensitive in detecting judgment differences than the Likert scale and other "equal-appearing interval scales" (288).

8.3.3 Perception Experiments Based On Indirect Judgments about Intonational Meanings

The most important type of perception experiment for the KIM is probably the "indirect identification task," also known as the "semantic task" in terms of Gussenhoven (1999). It was used to find and substantiate intonational categories since Kohler's (1987) seminal paper. The indirect identification test is a further development of the task that Nash and Mulac (1980) introduced. The basic idea behind this test is that a change of the intonation category causes a change in the meaning of the stimulus utterance. This meaning change and, hence, the existence of two separate intonation categories can be indirectly revealed by combining the stimuli with a constant context utterance, which is designed to match with the assumed meaning of only of the intonation categories. Then the listeners simply judge in the perception experiment whether the two presented utterances (stimulus and context) do or do not match. The principle of the indirect identification task is displayed in figure 8.2.

Recently the basic operation principle of the indirect identification test was refined by Landgraf (2014) and Niebuhr (2015) in the following way. Whenever an intonational function clashes with a coinciding lexical meaning, the corresponding utterance is interpreted as ironic (Bryant and Fox Tree 2005; Loevenbruck et al. 2013). So if listeners perceive a stimulus utterance as ironic in combination with intonation contour A, but not in combination with intonation contour B, then we can conclude that the listeners distinguished A from B and, moreover, associated contour A with a meaning that is diametrically opposed to the semantic content of the accented target word or phrase. This operation principle, which was called *irony task*, requires no constant context stimulus, which considerably lengthens the experiment and, due to its invariance, introduces a certain monotony into the task. The irony task has the further advantage that intonational meanings can be defined or tested in more detail, as they are the "negative images" of a specific word or phrase. In addition, it is reasonable to assume, and is supported by informal interviews with listeners, that "ironic" and "not ironic" are more common and intuitively applicable judgment categories than "matching" and "not matching" are.

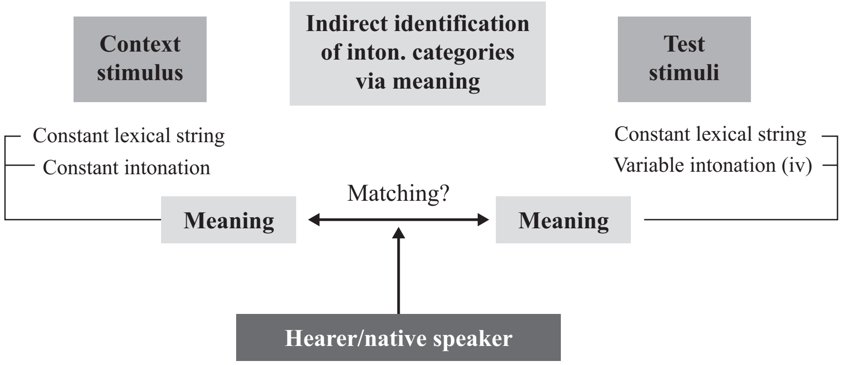


Figure 8.2
Schematic illustration of the makeup of indirect identification tests used to establish phonological categories in the KIM. inton, intonation.

8.3.4 Résumé

These descriptions of the KIM's methodological guideline and the corresponding perception experiments were to flesh out the fact that all phonological elements of the KIM (summarized in the section 8.4 below) are backed up by empirical evidence from at least one of these formal perception experiments. The phrase boundary categories are a small exception: the experiments did not rely on meaning-related judgments, but on similarly qualitative judgments about the degree of perceptual coherence of two adjacent stimulus utterances. As in semantic differential tasks, these judgments were made on a seven-step Likert scale.

Almost all phonological elements of the KIM have in the meantime been supported by consistent evidence from both direct and indirect judgment tasks. Perception experiments also served to identify and scrutinize trade-offs within and across distinctive features. For many categories or phonological contrasts, the evidence from perception was additionally cross-validated by acoustic analysis of production data obtained from specifically tailored read-speech tasks or the Kiel Corpus of Spontaneous Speech, or both. This applies in particular to pitch-accent peaks and valleys, concatenation contours, phrase boundary categories, and prominence levels, including classes and categories of emphasis.

8.4 Phonological Elements

8.4.1 Prominence Levels

Pitch accents are the elements of an utterance tune that stand out in terms of perceptual salience against the surrounding unaccented elements. This is a consensus of probably all intonation models, including the KIM. The KIM, however, goes one step further in that it assumes distinctive prominence levels and, hence, a separate phonological structure of accentual prominence that is orthogonal to the pitch-accent paradigms addressed in section 8.4.3.

The initial TTS-oriented version of the KIM distinguished nine prominence levels. This is surely a lot. But studies showed that listeners are able to project even more finely grained prominence scales onto perceived utterances (Fant and Kruckenberg 1989; Arnold, Wagner, and Möbius 2011). While prominence level 0 meant "unaccented," levels 1 to 7 represented different degrees of prominence. This basic framework still exists in the current version of the KIM. However, from the perspective of meaningful prominence differences, it turned out in perception experiments that a quadripartite accent phonology suffices (Kohler 1991a, 1991b). The corresponding PROLAB numbering is <0>, <1>, <2>, <3>.

Prominence level <0> still means "unaccented" and can thus not be associated with a pitch accent. In contrast, prominence levels <1> to <3> have to be associated with a pitch-accent category. Prominence level <1> refers to a secondary or "reduced" accent, level 2 represents a fully-fledged or "default" accent, and level <3> denotes an emphatically enhanced accent production. In the syntactic structure of PROLAB, prominence levels <1>, <2>, and <3> precede the corresponding pitch-accent label. Level <0> is not followed by any other PROLAB label. The communicative differences between the nonemphatic prominence levels <0> to <2> are illustrated in figure 8.3, based on an example of Kohler (1991a).

Figure 8.3 shows three realizations of *Er ist ins Kino gegangen* (He went to the cinema). The first utterance is realized with a fully fletched pitch accent on the initial syllable KI- of *Kino* (cinema), and a secondary/reduced pitch accent on the middle syllable -GAN- of *gegangen* (went). This sequence of prominence levels <2> and <1> on noun and verb is informationally neutral. That is, it puts the entire utterance content in focus. In contrast, reducing the secondary prominence on -GAN- further to <0> shifts the focus to the first prominence and in this way creates a (narrow-focus) contrastive interpretation of the utterance, for example, in the sense of, "He finally went to the cinema, although he actually wanted to go to the theater." Similarly, by increasing the prominence on -GAN- to the same level as in KI-, level <2>, the focus is shifted to the verb, so that the (narrow-focus) contrastive interpretation of the utterance changes, for example, to, "He finally walked to the cinema, although he actually wanted to go there by car."

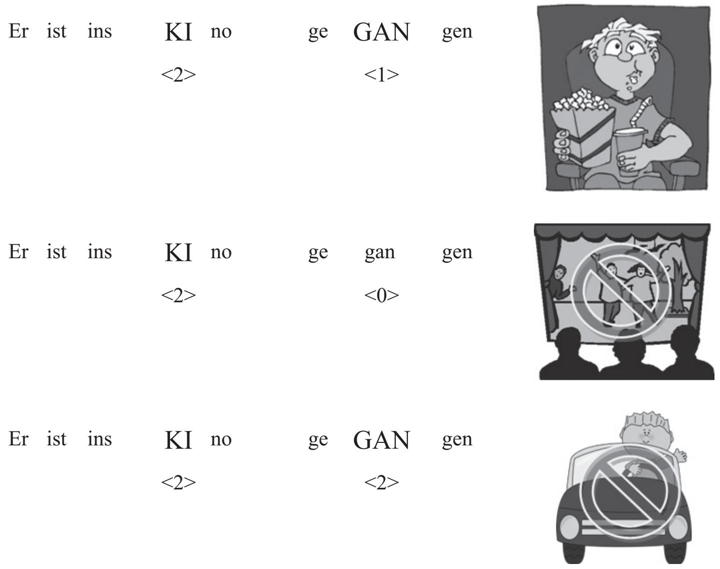

Er ist ins KI no ge GAN gen
<2> <1>
Er ist ins KI no ge gan gen
<2> <0>
Er ist ins KI no ge GAN gen
<2> <2>

Figure 8.3
Differences in utterance interpretation caused by prominence levels <0> to <2>.

A further increase from level <2> to level <3> on KI- or -GAN- maintains the respective narrow focus interpretation and additionally intensifies a certain aspect of the utterance (see section 8.4.6). For example, changing the level <2> accent on KI-no into an emphatic level <3> accent could signal, “Oh, how I envy him! I also wanted to go there today!” (positive intensification; see section 8.4.6).

In addition to these obvious meaningful effects of changes in relative prominence level, Kügler et al. (2015) recently showed that after a short introductory training, nonexperts are able to reliably apply the KIM’s phonological prominence-level distinctions to both spontaneous and read speech data. Interannotator agreement was even slightly higher in spontaneous speech, probably due to the fact that four prominence levels better reflect the complex prominence patterns of this speaking style than just two prominence levels (nonprominent versus prominent). According to feedback from annotators in Kiel, it is especially the availability of the secondary/reduced accent level <1> that makes it much easier to answer the often tricky question of whether a certain syllable is or is not pitch-accented. The corpus analysis of Niebuhr, Schroeder, and Baumann (2015) suggests that listeners make systematic use of all known prominence cues, that is, F0 movement, duration, and acoustic-energy level, when deciding between accent levels <1> and <2>.

8.4.2 Tune Structure and Phrase Boundaries

As illustrated in figure 8.4, a well-formed (i.e., noninterrupted) tune in the KIM consists of an intonation phrase with either four or five different types of syntagmatic slots, each of them having its own paradigmatic oppositions. A single-accent phrase includes a phrase-initial contour element, followed by a pitch accent, a phrase-final element, and a boundary category. When a phrase has more than one pitch accent, the two accents are linked by one of two phonologically different concatenation-contour categories. So unlike in many other intonation models (e.g., the autosegmental-metrical framework), the way in which pitch accents are linked is in the KIM not merely a phonetic consequence of the types of adjacent pitch accents or the temporal interval between them.

The original version of the KIM distinguished two types of phrase boundaries, which could roughly be termed “small/weak boundary” and “large/strong boundary.” Kohler (1997) already assumed that this simple binary distinction may be insufficient by admitting that “the linguistically and phonetically relevant categorization of these boundaries is not well understood” (10), and later analysis of spontaneous

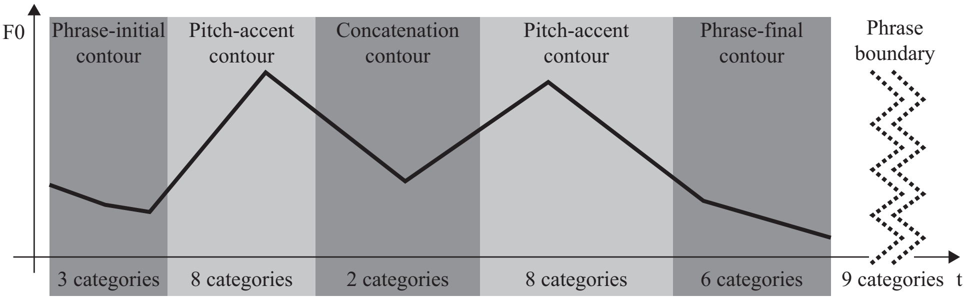


Figure 8.4
The syntagmatic structure of intonation phrases in the KIM. The concatenation contour occurs only in phrases with more than one pitch accent.

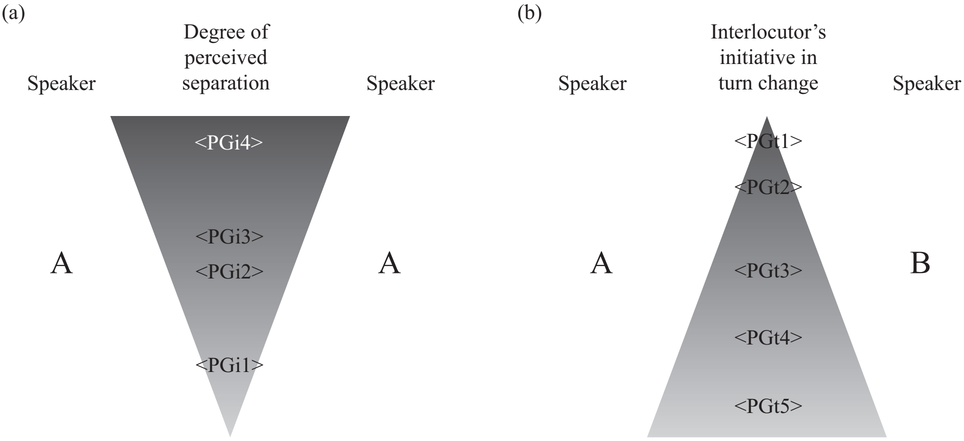


Figure 8.5
The phrase-boundary categories distinguished by the KIM for (a) turn-internal and (b) turn-final prosodic phrases.

speech dialogues and corresponding perception experiments by Peters (2012) led to a much more complex 4 + 5 system. The former number refers to turn-internal and the latter number to turn-final boundary categories. Figure 8.5 provides a summary of this system.

All turn-internal boundaries show similar degrees of phrase-final lengthening and intensity increases or decreases for rising or falling intonation movements, respectively. The distinguishing features are whether they are followed by a pause and end in an intonation movement that reaches into the limits of the speakers individual pitch range. Phrase boundaries without large final intonation movements and without following pauses are the weakest. They often occur unnoticed by speakers and listeners within syntactic constituents. Phrase boundaries without large final intonation movements but with a following pause are typically unavoidable "breathing pauses" within a speaker's line of argument. They can, but need not, occur at syntactic boundaries. A similar degree of perceptual separation is created by phrase boundaries with a large final intonation movement but without a following pause. They predominantly coincide with syntactic boundaries and are used, for example, for listing or contrasting arguments or items. Phrase boundaries that show both large final intonation movements and following pauses are the strongest and often are used to mark topic changes. They always occur at syntactic boundaries. The corresponding PROLAB labels are <PGi1>, <PGi2>, <PGi3>, and <PGi4> (<PG> stands for "Phrasen-Grenze" and <i> for "intern").

Turn-final phrase boundaries are also distinguished by whether the final intonation movement reaches into the limits of the speaker's individual pitch range. In addition, they take into account whether the phrase-final intonation is rising or falling and whether the turns end in overlap with the beginning of the interlocutor's turn. Together, these features reflect the speaker's and the interlocutor's relative contributions to the turn change (figure 8.5b). The two categories labeled <PGt1> and <PGt2> are the strongest turn-final phrase boundaries. They end in a large intonation movement and show no overlap with the interlocutor's turn. Both signal an intended turn-yielding, but <PGt1> ends in a fall, whereas <PGt2> ends in a rise, which in almost all cases is associated with a question. In contrast, <PGt3> and <PGt4> signal that the interlocutor can, but need not, take the turn. <PGt3> marks the end of a terminal low fall and is used when a speaker has completed her line of argument. The opposite is true for <PGt4>, which is why this phrase boundary can also occur in overlap with the beginning of the interlocutor's turn. Finally, <PGt5> indicates a forced turn-yielding and is characterized by a high final intensity level, no pronounced final F0 movement, and a phrase boundary in overlap with the interlocutor's turn.

8.4.3 Phrase-Initial Contour

The KIM distinguishes three types of phrase-initial contours, based on, for example, the experiments of Kohler (1991a, 2011):

- The *default phrase-initial contour* category rises from a lower pitch level to the level of the F0 maximum of first pitch accent. Its PROLAB symbol is <LP> (low prehead). It is a neutral way of opening a discussion or relating to previous utterances of the speaker or the interlocutor.
- *Falling phrase-initial contours* that start at a similar F0 level as the maximum of the first pitch accent are marked by <HP> (high prehead). Speakers use this phrase-initial category to convey their cooperativeness in the conversation. In addition to formal evidence from perception experiments, this attitudinal meaning is also supported by speech data analyses: <HP> occurs particularly often when speakers continue a line of argument or complete utterances of their dialogue partners.
- *High-falling phrase-initial contours* that set in and start falling at a much higher F0 level than the maximum of the first pitch accent represent a separate category in the KIM and are marked in PROLAB by <HP2>. Unlike <HP>, <HP2> is used to take up and critically revise what has been said before. In this sense, <HP2> conveys the opposite of cooperativeness. For example, in an utterance like "the car, the dog, and the children," an <HP> onset could be paraphrased as, "We have to take these things into account as well," whereas a <HP2> onset means, "This is what you forgot" or "This is why your plan won't work."

8.4.4 Pitch-Accent Contour

The KIM distinguishes two basic classes of pitch-accent contours: rising-falling peaks and (falling-)rising valleys (Niebuhr and Kohler 2004). Both contour classes have four distinctive intonation categories, which results in a total of eight different pitch accents.

The KIM does not make any restrictions concerning the occurrence of these eight categories in prenuclear and nuclear positions. Although Kiel Corpus analyses show that valley contours are more frequent in nuclear than in prenuclear positions, every accent category can basically occur with the same meaning in every position (see, e.g., Niebuhr and Zellers 2012 for pitch-accent peaks).

8.4.4.1 Pitch-accent peaks Synchronization plays a major role in distinguishing pitch accents in the KIM. This is most obvious for the peak categories, which are all defined by their synchronization. So unlike in, for example, autosegmental-metrical models, time or timing is directly phonologically relevant in the KIM instead of being an epiphenomenon of how and what tones are associated with the accented syllable. As for the latter, another important difference between autosegmental-metrical models and the KIM is that the Kiel model refers to the accented-vowel boundaries rather than the accented-syllable boundaries in describing differences between pitch-accent categories.

In its original version, the KIM distinguished three peak categories based on studies of Kohler (1987) and Niebuhr (2007c). However, in dealing with the Kiel Corpus of Spontaneous Speech, a fourth category was introduced, also inspired by the distinction between H* and L+H* in German GToBI (Grice and Baumann 2002). Kohler (2005) provided the perceptual evidence for this fourth category.

- The early peak reaches the F0 maximum before the accented-vowel onset. Thus, F0 is falling through the vowel to a low level. The PROLAB symbol for the early pitch accent is <)>. The early accent expresses that the corresponding piece of information is given and cannot or should not be changed. Depending on the semantic-pragmatic context, the specific meaning of early peaks can range from resignation ("there is nothing we can do about it") to reliability ("what I say is an immutable fact"). Not least for these reasons, the early category often occurs at the end of a speaker's turn.
- The medial peak reaches its F0 maximum within the accented-vowel boundaries, typically in the second half of the vowel. Thus, parts of the rising and falling slope are located inside the vowel. The corresponding PROLAB symbol is <^>. The medial peak is the neutral way of accentuation in German. It marks new pieces of information or is used to put a previously addressed piece of information up for discussion again (which is why even given information can occur with a medial peak; Niebuhr 2007c and Kohler 2005, 2007).
- The late-medial peak is characterized by a rise throughout the vowel. The rising slope starts before the accented-vowel onset and continues until after the accented-vowel offset. The F0 peak maximum is typically reached in the next sonorant sound segment. The late-medial peak is marked in PROLAB by <{>. It conveys that the corresponding piece of information is at odds with the speaker's expectation. The speaker makes "a matter-of-fact statement of this contrast" (Kohler 2005, 90).
- The late peak starts rising after the accented-vowel onset and continues rising until after the accented-vowel offset. The peak maximum is located quite high in the speaker's pitch range and typically reached still later than for the late-medial peak, that is, in the vowel of the following unaccented syllable. However, unlike for the late-medial peak, the rise must immediately change into a steep fall, part of which has to be realized within the accented word (Niebuhr and Zellers 2012). The PROLAB symbol of the late peak is <(>. The late peak also signals a contrast to the speaker's expectation, but in addition to the late-medial peak, the late peak "adds a personal expressive evaluation to the contrast" (Kohler 2005, 90). Depending on the semantic-pragmatic context, this can signal either a positive surprise or incredulity in the sense that the speaker is not willing to simply accept the facts that led to this contrast.

To make the meanings and meaning differences of the four pitch-accent peaks more tangible, we can combine them with the simple particle "no." In this case, their meanings can be paraphrased as follows: "And that's that" (early), "I'm against it, but let's discuss that" (medial), "This comes unexpected, but I can live with that; I can't change it anyway" (late-medial), and "I can't believe that this is happening to me, and I will do something about it" (late).

Figure 8.6a illustrates the synchronization characteristics of the four pitch-accent peaks relative to the accented-vowel boundaries. More recently, it was found that peak

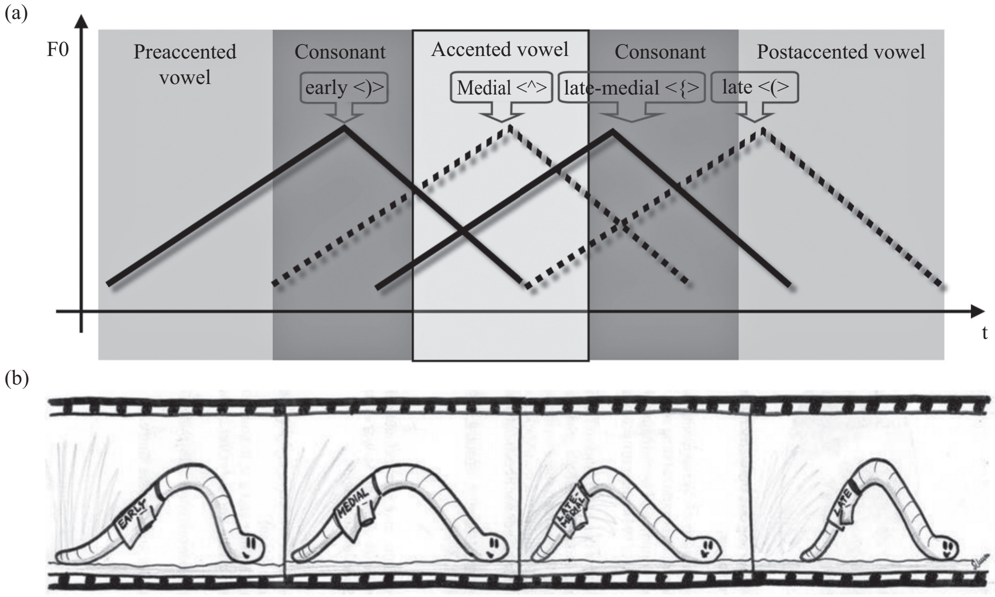


Figure 8.6
The four pitch-accent peaks of the KIM shown (a) in their distinctive synchronization windows and (b) in additional combination with their characteristic peak shapes, which together constitute the "worm paradigm." *Source*: Drawing courtesy of Nathalie Schümchen.

shape also matters, particularly with respect to the steepness of rising and falling slopes. Some shapes are more effective than others in signaling pitch-accent categories, and as a consequence, their synchronization ranges become larger and overlap with those of other categories across the vowel boundaries. For example, a peak with a slowly rising-falling shape is still perceived as being medial, even when the F0 maximum is located well after the accented-vowel offset. A peak with a fast falling shape is still perceived as being early, even when its F0 maximum is clearly reached inside the accented vowel (Niebuhr 2007a, 2007c). In contrast, accent peaks with a plateau-shaped maximum need not rise at all into the accented vowel to be identified as being medial (Niebuhr 2011). This perceptually determined interplay of peak synchronization and peak shape also manifests itself in production, for example, in the form of speaker-specific trade-offs between the two parameters, with "shapers" and "aligners" at the extreme ends of the trade-off continuum (Niebuhr et al. 2011).

When the four pitch-accent categories are depicted in their defined synchronization window and with their most effective peak shape, they look like still images of a worm that successively crawls across the accented vowel (see figure 8.6b). Therefore, the four accent peaks together form a set of contrasts that is also called the "worm paradigm." The worm starts crawling before the accented vowel with a slowly rising, fast-falling shape (early), takes a more symmetrical slowly rising-falling shape inside the vowel (medial), crosses the vowel offset with a fast-rising slow-falling shape (late-medial), and ends his trip through the accented vowel in the post-accented vowel, having a symmetrical fast rising-falling shape (late).

8.4.4.2 Pitch-accent valleys The attitudinal meanings of peak categories are oriented toward a particular piece of information or toward the speakers themselves. In contrast, valleys bring in the addressee. While the meanings of peaks and valleys point broadly in diametrically opposed directions, Kohler stresses that from a phonetic-phonological perspective, “valleys are not the mirror images of peaks, and it is not surprising that they should have different perceptual status” (1991a, 308). In terms of phonological contrasts, this means, for example, that there is only perceptual evidence for a binary distinction based on synchronization (Kohler 1991a; Niebuhr and Kohler 2004; see figure 8.7a).

- Early valleys reach their F0 minimum before the accented-vowel onset. Thus, F0 rises throughout the accented vowel. The corresponding symbol in PROLAB is <]>. Early valleys invite dialogue partners to make a matter-of-fact contribution to the discussion, for example, by answering a question or expressing their opinion about a certain piece of information.
- Late valleys differ from early valleys in that they start rising inside the accented vowel. The PROLAB symbol of late valleys is <[>. Like early valleys, late valleys also extend an invitation to the dialogue partner to contribute to the discussion. However, in the case of late valleys, this invitation is aimed at the interlocutor’s feelings about a certain piece of information.

So a discourse marker such as “okay” can call for backchanneling, agreement, or disagreement in combination with an early valley, but means, “How do you feel about it?” or, “Can you live with that?” in combination with a late peak.

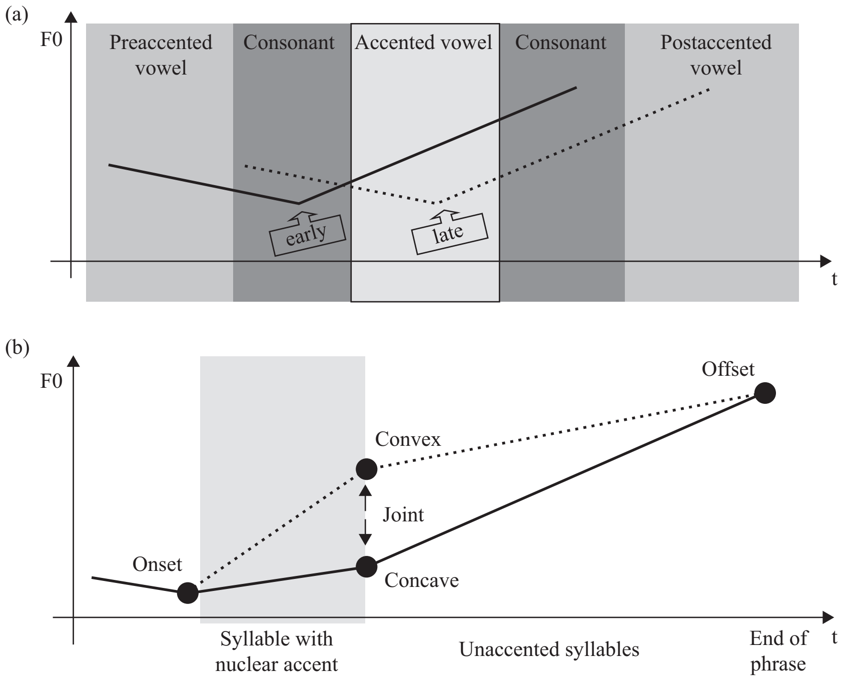


Figure 8.7
The two pitch accent valleys of the KIM shown (a) in their distinctive synchronization windows and (b) in combination with their distinctive feature of a convex or concave rise.

In addition to the phonological distinction based on synchronization, a further distinction is based on shape. This is another aspect in which valleys differ from peaks, for which shape is (as far as we know today) not an independent phonological feature. However, in the case of valleys, F0 rises can take a convex or concave shape (seen from above), for example, based on a hypothetical joint located at the end of the accented syllable (see figure 8.7b):

- Valleys with a convex shape "restrict" dialogue partners in the sense that they reduce their alternatives in the conversation.
- Valleys with a concave shape "activate" dialogue partners and invite them to shape the further discussion at their own discretion.

In other words, from the speaker's point of view, with the convex shape, the speaker remains "at the helm," whereas, with a concave shape, he or she hands the control over to the interlocutor. Depending on the semantic-pragmatic context, this can result in different interpretations. For example, in statements like "Then came Anna," concave shapes ask for conformation or a short comment, whereas convex shapes signal "hold on and let me continue." In the case of questions like "Are you Anna?," the meanings of convex and concave shapes can be paraphrased as "I just ask for your name and for nothing else; I continue speaking afterward" (convex) or "Please also tell me a bit more about you" (concave). Correspondingly, convex valleys occur more often turn-internally and concave valleys turn-finally.

Production and perception experiments by Dombrowski and Niebuhr (2005, 2010) put the convex-concave distinction on solid empirical grounds for the early valley category (cf. figure 8.4b). However, there is also initial evidence from pilot perception experiments that the distinction also applies to the late valley category. As the PROLAB inventory was last updated in 2010, there are no symbols for the convex-concave distinction as yet. However, an obvious solution would be to add an <x> for convex and a <c> for concave right before the respective synchronization label. Hence, <2x]> would, for instance, denote a convex early valley pitch accent at a default prominence level.

8.4.4.3 The minus accent Phenomenologically similar to the concept of phrase accents in GToBI (Grice, Ladd, and Arvaniti 2000) but syntagmatically not restricted to post-nuclear prominences, the KIM takes into account that syllables can be made prominent for rhythmical purposes only. Barry (1981) points out the important guiding function of speech rhythm (in West Germanic languages). If syllables are made prominent for rhythmical purposes, they need not be associated with a pitch accent, that is, with a separate attitudinal meaning. Therefore, the KIM offers the possibility of marking prominent syllables without a pitch-accent label. In this case, the phonological prominence level—often <1>, sometimes <2>—is followed by a minus sign <–>, which is why the F0 course on the corresponding syllable is also called the *minus accent*. Analyzing the annotations in the Kiel Corpus of Spontaneous Speech, Peters, Kohler, and Wesener (2005) found that minus accents are relatively rare, but not exotic (about 7 percent of all prominent syllables). They occur typically in phrase-initial or phrase-final position. In the latter position, they are often related to disfluencies. About 20 percent of all minus accents represent prominent syllables inside "hat patterns" (see section 8.4.5). Not least for this reason, minus accents are typically annotated when F0 runs flat across a prominent syllable.

8.4.5 Concatenation Contour

The separate phonological categorization of the F0 course in between two accented syllables is probably one of the two most noticeable features in which the KIM differs from other models of intonation, next to the KIM's prominence-level distinction. In the original version of the KIM (Kohler 1991a, 1991b), there was a tripartite distinction among the concatenation contours, symbolized in PROLAB by <0.>, <1.>, and <2.> (see figure 8.8a).

The <0.> category means that there is no F0 indentation between two pitch accents, which results in a so-called *hat pattern.* In contrast, <1.> refers to a moderate and <2.> to a strong F0 indentation, the latter reaching down to the bottom of the speaker's individual pitch range. However, a more detailed and comprehensive series of perception experiments summarized in Ambrazaitis and Niebuhr (2008, 2014) was inconsistent with this phonological analysis. First, the experiments showed that hat patterns do not have to be completely flat. Even a slight F0 indentation of about two or three semitones can still have the same communicative function as a completely flat hat. Second, Ambrazaitis and Niebuhr found no empirical support for a distinction between <1.> and <2.>. As is displayed in figure 8.8b, this led to a revision of the concatenation-contour paradigm in favor of a binary phonological distinction: hat pattern versus dip pattern. It is suggested here that the former is marked by <h> and the latter by <d>.

The hat pattern looks like a staple for a stapler, and in fact it has a bracketing function that holds two or more accented pieces of information together, whereas the opposite applies to the dip pattern. So, for example, when produced with the hat pattern, the greeting "Dear ladies and gentlemen" is content-wise the same as "Dear audience," whereas a dip pattern signals that the two groups of listeners are addressed

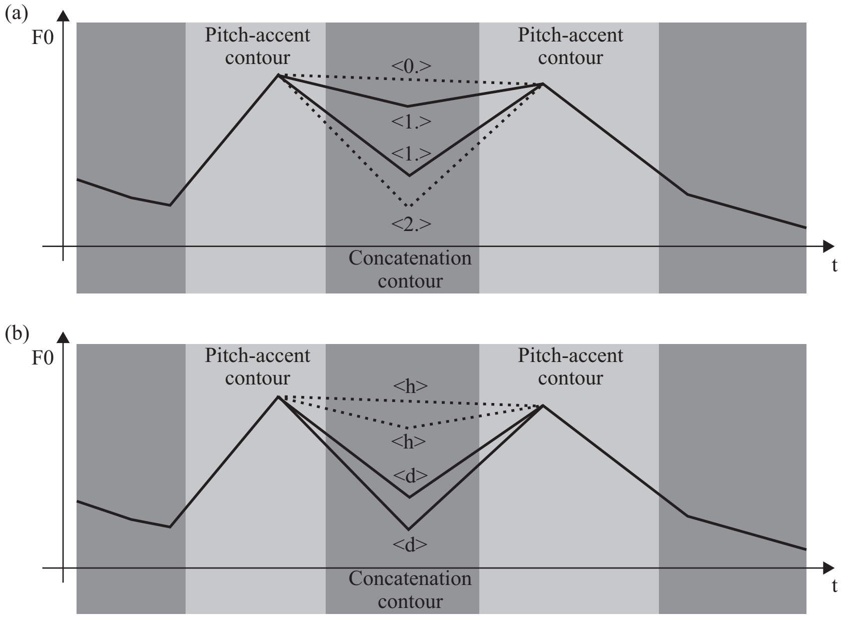


Figure 8.8
The concatenation contours of the KIM. (a) The original paradigm with three categories. (b) The revised paradigm that distinguishes only two categories: hat and dip contours.

separately (as if they were sitting on opposite sides of the room). Likewise, “Would you like tea or coffee?” with a hat pattern asks the interlocutor to choose between a closed set of two alternatives, whereas the same question with a dip pattern becomes an open list of suggestions, equal to “Would you like something to drink?” Also, interlocutors are basically given the option to ask for a hot chocolate, which would not be possible in connection with the hat pattern.

In the context of concatenation phenomena, it must also be noted that the KIM distinguishes between upstep and downstep (Kohler 1991a, 1991b, 1997). This applies to pitch-accent peaks. If the F0 maximum of pitch-accent peak B is higher than that of the previous peak A, then peak B is marked as upstepped by placing the PROLAB label <|> in front of the peak-category label. However, upstep marking requires, per definition, that the two accent peaks have the same phonological prominence level. So <2^ d |2^> is a valid upstep annotation, but <1^ d |2^> is not. This restriction is to avoid redundancies, as higher prominence levels are associated with larger F0 peak ranges anyway.

8.4.6 Phrase-Final Contour

Given the contour differences between pitch-accent peaks and valleys, it is logical that each of these two contour classes has its own phonological paradigm of phrase-final contours (the two paradigms are depicted in figure 8.9a and 8.9b):

- Peak contours can end in a terminal fall until the lower limit of the speaker’s individual pitch range. This is marked in PROLAB by <2.>. The categoricalness, dominance, and potential impoliteness conveyed a terminal fall can be attenuated “without appearing to be indecisive and insecure” (Ambrazaitis 2005, 198). This is done by attaching a small, hardly audible rise of about one to two semitones (five to fifteen hertz; Peters 2000) to the terminal fall. The attenuated terminal fall was called “pseudo-terminal” and is symbolized in PROLAB by <2;>. Peters (2000) discovered by analyzing <2;> annotations in the Kiel Corpus of Spontaneous Speech that pseudo-terminal falls are far more often produced by women than by men, probably due to the traditional role model of women in a patriarchal society.
- Falls not reaching into the lower limit of the speaker’s pitch range are called “nonterminal.” This type of phrase-final peak contour, which is marked by <1.> in PROLAB, typically occurs when speakers reflect on different alternatives. For example, they make a statement, and then they consider it from a different perspective in the subsequent intonation phrase. A prime example are statements such as “We can do it” or “A great idea” at the end of which the interlocutor already hears (due to <1.>) that a “but” will follow.
- Cases in which a peak contour changes at the F0 maximum into a level plateau are indicated by <0.> in PROLAB. Phrase-final <0.> contours typically occur in turn-medial position. So like <1.>, <0.> signals that the speaker will continue. But unlike <1.>, <0.> is used when a statement is continued or further justified rather than contrasted with a different statement in the next phrase.
- Phrase-final contours after valley contours are distinguished by how high they rise. There are two phonological categories of phrase-final rises: a medium rise and a high rise. The original version of the KIM (Kohler 1991a, 1991b) did not specify how exactly the two rise categories are defined. Recent evidence from production and perception (see also Michalsky 2014) led to the following definitions (see figure 8.6b): medium rises do not reach beyond the F0 maximum of the immediately

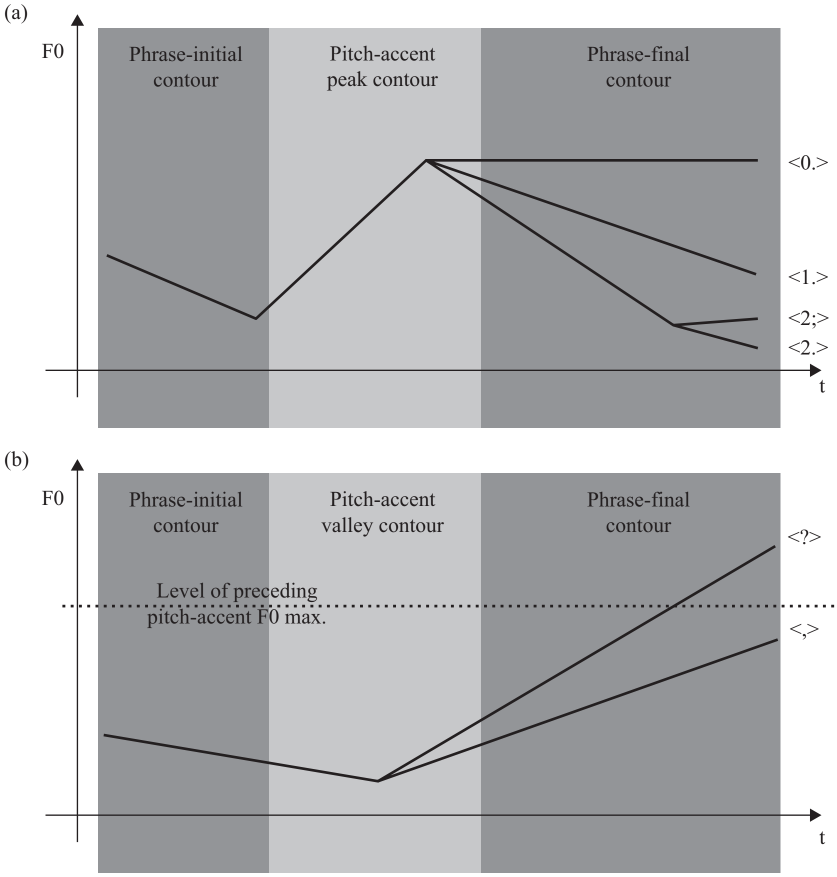

(a)
F0
Phrase-initial contour
Pitch-accent peak contour
Phrase-final contour
<0.>
<1.>
<2;>
<2.>
t
(b)
F0
Phrase-initial contour
Pitch-accent valley contour
Phrase-final contour
<?>
Level of preceding pitch-accent F0 max.
<,>
t

Figure 8.9
The paradigmatic oppositions among the phrase-final contours of the KIM after (a) pitch-accents peaks (four categories) and (b) pitch-accent valleys (two categories).

preceding pitch-accent contour. The opposite applies to high phrase-final rises. That is, they continue beyond the F0 maximum of the immediately preceding pitch-accent contour. Medium phrase-final medium rises are symbolized by <,> and have a similar "continuation function" as <1.> and <0.> after peak contours. However, unlike the latter two categories, the <,> category is listener-oriented rather than information-oriented.

- High phrase-final rises are the counterpart of phrase-final <2.> falls. Thus, they are also called "terminal rises" (Kohler 1991a, 1991b). Their PROLAB symbol is <?>. As is already indicated by the question-mark symbol, high phrase-final rises typically occur at the ends of interrogative utterances or when the turn is for other reasons handed over to the interlocutor.

8.4.7 Additional Categories of Emphatic Intonation

Work with the Kiel Corpus of Spontaneous Speech over many years made it clear that emphatic expressions are anything but rare in informal everyday conversation. Therefore, the KIM started a new line of research around 2006, aimed at discovering the form-function links of emphatic intonation contours in German and consolidating them into a system of phonological paradigms that supplements the KIM's regular intonational phonology.

Emphasis is "a very broad and general term applied to any kind of prominence attached to a linguistic element" (Trask 2004, 188). Newman (2015, 73) calls it "a convenient phonological catch-all." The use of the term *emphasis* in the literature ranges from function-oriented definitions such as "emotional involvement on the part of the speaker" (Ladd 1996, 129) to signal-oriented definitions in which emphasis is an acoustic feature reflecting the "relative energy in the higher frequency bands" (Heldner 2003, 40). Against this complex terminological background, the KIM provides an explicit definition of emphasis. This definition combines form and function and separates emphasis from the noncontrolled (i.e., physiologically determined) long-term effects of emotion on speech. Emphasis manifests itself as local extra effort, added by the speaker to both articulatory and phonatory patterns with the intention to intensify an aspect of the communication.

For these intensified aspects of communication, figure 8.10 shows that the KIM makes a tripartite distinction, inspired by the communication models of Bühler (1934) and von Thun (1981), as well as the notion of emphasis of British school phoneticians (Jones 1969):

- *Emphasis for intensity* is an accent-level phenomenon that intensifies the semantics of the accented piece of information. Emphatic accents are not just regular pitch accents produced with more articulatory and phonatory effort. Rather, this extra effort adds an expressive evaluation to the meaning of the pitch accent.
- *Emphasis for contact* relates to the intonational nucleus. It intensifies the speaker's connection to the interlocutor—for example, to assess this connection or bind the two dialogue partners closer together.
- *Emphasis for attention* is a phrase-level phenomenon whose intensification is directed toward the communication channel, making the interlocutor more receptive or susceptible to the speaker's message.

A fourth potential category would be *emphasis for contrast*, which relates to the informational or situative context of the communication (Kohler 2006). It includes the late-medial and late pitch-accent categories and the corresponding patterns of

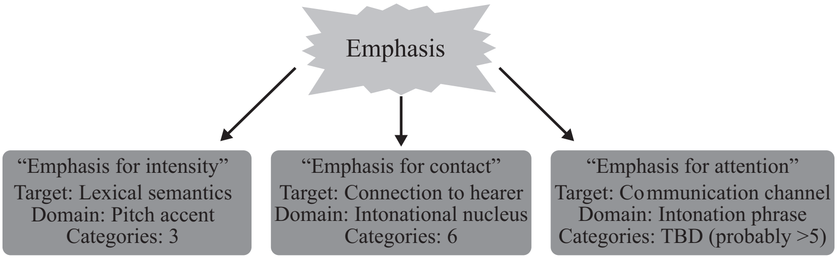

Emphasis
"Emphasis for intensity"
Target: Lexical semantics
Domain: Pitch accent
Categories: 3
"Emphasis for contact"
Target: Connection to hearer
Domain: Intonational nucleus
Categories: 6
"Emphasis for attention"
Target: Communication channel
Domain: Intonation phrase
Categories: TBD (probably >5)

Figure 8.10
The three emphasis classes distinguished in the KIM. Each class includes at least three phonological categories. The exact number of categories is still unknown as the emphasis system is subject of ongoing research. TBD, to be determined.

phonological prominence. It is not entirely clear yet whether these accent patterns meet all criteria of the KIM's emphasis definition (e.g., extra effort in the articulatory and phonatory domain; Görs and Niebuhr 2012). For this reason, "emphasis for contrast" was omitted in figure 8.10.

Overall, evidence from combined production and perception studies have identified more than a dozen emphasis categories. Providing detailed portraits for all of these categories would make this chapter far too long. Therefore, the following sections just give a brief overview of each class of emphasis categories.

8.4.7.1 Emphasis for intensity Emphasis for intensity is the oldest and best investigated class of emphasis phenomena within the KIM. Studies by Kohler (2006) and Niebuhr (2010) suggested distinguishing three categories: positive intensification, negative intensification, and reinforcement. The corresponding acoustic profiles are sketched in figure 8.11.

Positive intensification is phonetically based on the strengthening of sonorous features in the accented syllable. For instance, the accented vowel is strongly lengthened, partly at the expense of the preceding consonant. The F0 rise into the vowel freezes at a high level and creates an extensive plateau that is left around the end of the vowel in a shallow fall. The acoustic energy contour also increases and decreases slowly at the onsets and offsets of the vowel, in this way paralleling the F0 deceleration. In contrast, negative intensification weakens sonorous features. This involves a quickly articulated accented vowel with large and abrupt acoustic energy changes at both edges and a pointed F0 peak that falls steeply to a low level right after the accented-vowel onset. A pressed and irregular voice quality adds to the "barking" impression given by the vowel. Unlike the vowel, the preceding consonant is strongly lengthened. Reinforcement is a kind of hybrid form that shares features of positive and negative intensification. Reinforcement resembles negative intensification in terms of the lengthened consonant at the accented-syllable onset and the high F0 and acoustic energy dynamics inside the accented vowel. However, as for positive intensification, F0 is rising into the vowel; the vowel is not shortened and shows a modal voice quality (Landgraf 2014).

Emphasis for intensity underscores the semantics of the accented word. But while positive and negative intensification target the word's valence, reinforcement is aimed

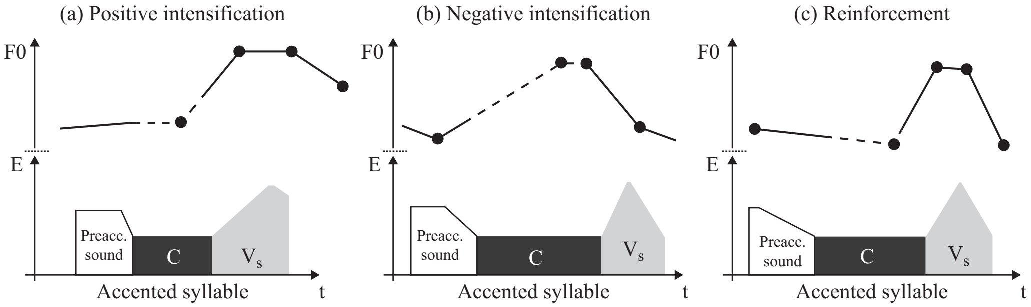


Figure 8.11
Prosodic profiles of the three emphatic categories in the KIM that aim at intensifying lexical semantics, particularly of adjectives: (a) positive intensification, (b) negative intensification, and (c) reinforcement. Preacc = preaccented.

at the truthfulness of the information represented by the accented word. Depending on the emphasis category, a sentence such as "This was a HUGE traffic jam" can additionally convey the following: "That's awesome. I have never seen so many cars before in one place!" (positive intensification); "Oh, crap; this will take forever!" (negative intensification); or "You can trust my assessment. I have seen a lot of traffic jams in life!" (reinforcement). Corpus analyses showed that emphasis for intensity is not restricted to adjectives like *HUGE*, although this is the most frequently used word class. Moreover, all three emphasis categories basically occur with all categories of pitch accent. Still, early peaks and nonearly peaks are more frequently associated with negative or non-negative intensification, respectively.

In PROLAB, emphasis categories are represented by capital letters placed between the phonological prominence level (typically <3>) and the pitch-accent category. In the case of emphasis for intensity, the PROLAB labels are <P> for positive intensification, <N> for negative intensification, and <R> for reinforcement.

8.4.7.2 Emphasis for contact One type of emphasis for contact refers to the pairs of intonational plateaus that are also known as "calling contours" (henceforth *nuclear one-step down contours*). Production and perception studies showed that there are actually three different categories of such nuclear one-step-down contours: they express reluctance, harmony, or disharmony (Niebuhr 2015).

One-step-down contours signaling reluctance are characterized by a medium-sized rise to the first plateau, followed by a tiny downward step of about one or two semitones to the second plateau. The second plateau is much longer than the first. Both plateaus have similarly high acoustic energy levels and start similarly "late" in the vowels of the accented and postaccented syllables. The plateau sequence expressing harmony is also characterized by a medium-sized rise up to the first plateau, but the step down to the second plateau is about twice as large as in the reluctance condition, that is, between three and four semitones. Both plateau durations and acoustic energy levels decrease from first to second plateau. Harmony plateaus start almost simultaneously with the respective vowel onset. Disharmony differs from harmony in that duration and acoustic energy levels increase rather than decrease from first to second plateau. Moreover, compared with harmony, disharmony is associated with larger upward and downward pitch steps to the first and second plateaus. The plateau alignment is intermediate between those of harmony and reluctance.

A second type of emphasis for contact adds another rise to a nuclear peak contour, in this way creating a rising-falling-rising intonational nucleus. Functionally, these nuclear contours are manipulative in the sense that they infringe on the interlocutor's self-determination. This can be done for three different purposes that have clearly distinct prosodic profiles and are thus analyzed as different emphasis categories: threatening, pleading, and confessing. *Threatening* is prosodically characterized by fast and extensive F0, duration, and intensity changes. In brief, syllables with rising or high pitch are long and loud; syllables with falling or low pitch are short and soft. *Pleading* is based on strong lengthening of the accented syllable, a very high reaching phrase-final rise, and a constantly high acoustic energy level. *Confessing* involves a strong lengthening of unaccented syllables and a reduced F0 range at a constantly high level.

Given that emphasis categories are indicated in PROLAB by a capital letter in between accent level and pitch-accent category, an obvious decision was to label the three one-step contours by <W> for reluctance ("<W>iderwille" in German), <H> for harmony, and <D> for disharmony. The three rise-fall-rise contours receive the letters

<T> for threatening, <B> for pleading ("<B>itten" in German), and <C> for confessing. So possible PROLAB annotations would be <3W^> and <3T(>, for example.

8.4.7.3 Emphasis for attention The class of emphasis-for-attention categories is the least well investigated in the KIM. At the current state-of-the-art, it includes two pairs of categories based on the reiteration of either words (e.g., "very very nice!" or "go go!") or pitch-accent patterns (produced in a syllable-by-syllable or foot-by-foot fashion, as in "eve-ry-sin-gle-day!" (see the chanting concept in Auer, Couper-Kuhlen, and Müller 1999). These reiterations are integrated into an overarching, coherent prosodic structure (e.g., Niebuhr et al. 2012). An additional type of phrase-level phenomenon subsumed under emphasis for attention is what Strangert (2003) called "emphasis by pausing." Common to all types and categories of emphasis for attention is that the extra effort is spent on disrupting the fluency or rhythmicity of an utterance. However, as the research on the forms and functions of all these phenomena is still ongoing, it would be premature to specify the exact number of categories and suggest corresponding PROLAB labels here.

8.5 Working with the KIM

Researchers who are interested in working with the KIM can find all relevant details about the syntactic rules of PROLAB in the KIM-Trainer compiled by Peters and Kohler (2004). The trainer also provides instructive examples and practical exercises, and it summarizes previous experiences with PROLAB annotations. Only some very basic points of the trainer should be mentioned here.

8.5.1 The Annotation Cycle

As is suggested by Kügler et al. (2015), annotators should start by determining the location and categories of phrase boundaries. Then the corresponding speech section should be listened to several times to define its prominence structure in terms of the four prominence levels. Each lexically stressed syllable of a word is assigned a number between <0> and <3>. The labels are set at the vowel onsets of the stressed syllables. In the following step, pitch-accent categories should be determined for all stressed syllables with prominence levels above 0. The labels of accent categories are put after the prominence number. In the case of an upstepped pitch accent, <|> is added before the prominence level in this step of the annotation. The final step of the annotation fills the gaps between the pitch accents. That is, phrase-initial, phrase-final, and concatenation contour categories are identified, and the corresponding PROLAB labels are added. The phrase-initial and -final labels are set at the beginning or end of the prosodic phrase, respectively, so phrase-initial/-final contour labels and phrase-boundary labels share the same time stamp. Labels of concatenation contours are placed at the same point in time as the pitch accent at which the concatenation contour ends (the second of two pitch accents or the rightmost one in a sequence of three or more pitch accents).

8.5.2 Potential Problems

Although it was found that the prominence level distinctions offered by the KIM can be reliably annotated, prominence levels are sometimes hard to specify. To solve this problem, it proved helpful to first listen to a longer stretch of the speech material to be annotated. This gives the annotator an impression of how the reduced prominence level <1> sounds in comparison to levels <0> and <2> for that inidividual speaker. The same also applies to hat patterns in comparison to dip patterns and pseudo-terminal in comparison to terminal and nonterminal phrase-final falls.

Another tricky distinction is that between peak and valley contours at the beginning of a hat pattern. The preliminary convention is to label high-reaching rises, particularly those with a concave shape, as valley contours and all other hat-pattern initial rises as peak contours.

In general, it is highly recommended that annotators familiarize themselves with the categories of the model and also learn to realize them in contrast to one another. On this basis, annotators can imitate what the speaker produced and, moreover, contrast the intonation category in question with its paradigmatic neighbors. Such ad hoc contrasts often facilitate the decision-making process in prosodic annotation. However, learning to produce all intonation categories and contrasts requires considerable supervised training and can easily take several months.

8.6 Perspectives for Further Developments

This section briefly summarizes findings that have been obtained within the KIM framework but which are only insufficiently or not at all taken into account by the current version of the model. The findings come from three different fields of research.

8.6.1 Vowel Boundaries Are Not Directly Relevant

A key assumption of the KIM is that the peak and valley pitch-accent contours are distinguished by the synchronization of contour points relative to onset and/or offset of the accented vowel. However, a growing body of evidence undermines this vowel-related phonological concept. For example, for a stepwise peak shift across the C-V boundary of the accented syllable, pitch-accent identification changes sooner from early to medial, when the intensity difference between consonant (C) and accented vowel (V) is smaller, that is, when the consonant already has almost the same intensity level as the following vowel. Moreover, the identification change from early to medial is more abrupt when the intensity increase into the vowel is faster. Analogous effects of magnitude and slope of the intensity change on pitch-accent identification were found for a stepwise peak shift across the V-C boundary of the accented syllable and the corresponding change from medial to late pitch-accent category (Niebuhr 2006, 2007a).

These findings indicate that it is not the vowel boundary in terms of the change in formant pattern that matters in the synchronization and identification of pitch accents. Rather, what matters are the increases and decreases in intensity caused by the accented vowel and how the pitch-accent contour is positioned (and even shaped) relative to the resulting areas of higher and lower intensity. Accordingly, all other things being equal, different accented-syllable compositions lead to different identification boundaries of pitch-accent categories within the same peak-shift continuum. This fact is not least reflected in the similar effects of syllable structure on pitch-accent alignment that were measured across many languages. Shifting the focus in synchronization from sound segments to their intensity patterns also takes into account that pitch-accent categories can also be realized and distinguished on interjections like “mmm,” which have no separate vowel segment. Speakers produce the prolonged nasal [m:] in “mmm” with a rising-falling intensity pattern, and shape and synchronize the pitch-accent contour relative to this intensity pattern, just as in a CVC syllable (Kirsch 2015).

8.6.2 Intonation beyond F0, Part 1: Pitch-Accent Specific Micro-Rhythms

Pitch accents are not simply put on top of a string of sound segments. Rather, they change the durations and intensity levels in the environment in which they occur, that is, the triplet of preaccented, accented, and postaccented syllable. Pitch accents create an imprint on this triplet in the form of a pitch-accent–specific micro-rhythm. Compared with a medial pitch accent, for which the preaccented and postaccented syllables are rhythmically symmetrical, that is, they have about the same low duration and intensity levels and hence appear similarly nonprominent, the early pitch accent slightly strengthens the preaccented syllable by making it longer and louder and further weakens the postaccented syllable by making it even shorter and softer. The opposite applies to the late pitch accent.

These prominence patterns and the resulting micro-rhythms in the triplet of preaccented, accented, and postaccented syllable also emerged in a finger-tapping task: speakers are asked to produce short test utterances with early, medial, and late peaks while tapping their index finger for each syllable on a device that records length and strength of these finger-tappings (Niebuhr 2019). Niebuhr further showed, in line with Niebuhr and Pfitzinger (2010), that listeners can identify early, medial, and late pitch accents only by hearing their corresponding micro-rhythms. Moreover, combining pitch accent peaks in early, medial, and late synchronization with mismatching micro-rhythms (in resynthesized stimuli), diminishes pitch-accent identification and makes the corresponding stimulus utterances sound less natural.

8.6.3 Intonation beyond F0, Part 2: Segmental Intonation

Voiceless fricative sounds, including postaspiration sections, all convey aperiodic pitch impressions. A number of acoustic analyses across languages and language varieties showed that the spectral patterns of such fricative sounds do not remain constant in different F0 contexts. Rather, the spectral energy distribution of a fricative sound varies such that the aperiodic pitch impression it creates changes in parallel to the adjacent F0 context (Niebuhr 2008, 2009, 2012; Ritter and Röttger 2014; Zygis et al. 2014). Projected onto English, this would mean, for example, that the /ʃ/ in the middle of *mushroom* sounds higher in the context of a high pitch-accent peak and lower in the context of a low pitch-accent valley. Likewise, the post-aspiration at the end of *lack* would sound higher in the context of a phrase-final F0 rise but lower in the context of a phrase-final F0 fall.

In addition, perception experiments provided evidence that listeners can integrate context-adequate aperiodic pitch impressions in their perception of intonation contours, for example, in that they fill in utterance-internal F0 gaps or restore truncated phrase-final F0 movements (Kohler 2011; Mixdorff and Niebuhr 2013). Particularly with respect to this potential perceptual relevance, F0 or pitch-oriented fricative variation in speech was termed *segmental intonation*.

8.6.4 Implications for the Further Development of the KIM

In order to cope with the new empirical findings in sections 8.6.1 to 8.6.3, the KIM must consistently implement the principles on which it is based. The current KIM is inconsistent insofar as it sets up phonological categories based on groups of listeners and the intonational meanings they identify. But it then defines these phonological categories by acoustic rather than perceptual features. Furthermore, the KIM takes an intermediate position in the conceptual issue of contour and tonal-target (level) models. It places contours in the center of the phonology, but at the same time decomposes these contours into single-movement features or local F0 landmarks, for example, in the case of pitch-accent and concatenation categories. However, this "pattern approach" is not pursued as rigorously as it could be.

So, in the opinion of this author, the further development of the KIM requires two steps:

1. With respect to the model's *paradigmatic* structure, the focus in defining intonational contrasts must be shifted from acoustics to perception.
2. With respect to the model's *syntagmatic* structure, the focus in defining the building blocks of intonation must be shifted from a heterogeneous mixture of contours and local F0 landmarks to Gestalt-like patterns of perceived pitch.

An initial basis for this further development was laid out by Niebuhr (2007b) in the *contrast theory of intonation perception* (C-TIP), see also Barnes et al. (2020). The theory capitalizes on the fact that perception follows the same principles across modalities. Therefore, the construction of intonational categories can be conceived in analogy to the relatively well-understood construction of visual objects. A basic assumption of the contrast theory is that intonational categories consist of two gestalt-like patterns: a pitch gestalt and a prominence gestalt that is associated with the pitch gestalt.

Pitch gestalts are constituted by tonal elements. According to the "spectral constraint" and "tonal movement coding" principles of House (1990, 1996), the F0 course is perceptually broken down by the listeners into a sequence of pitch levels and pitch movements. Pitch levels are directly translated into single tonal elements; pitch movements are represented by two tonal elements, one at each end of the movement. Aperiodic pitch impressions and their respective tonal elements are integrated in the pitch gestalt. The number of perceived movements and level tones and hence the total number of tonal elements within a gestalt can vary (e.g., due to the duration of F0 slopes and the coinciding segment/formant structures), but this variation is phonologically irrelevant as long as the gestalt pattern as a whole remains constant.

In the case of the KIM's three oldest pitch-accent categories (early, medial, and late peak), the pitch gestalt is always the same: rising-falling. Figure 8.12a illustrates the simplest rising-falling pattern, based on three tonal elements. It also shows that the phonological differences among early, medial, and late peaks lie in the prominence gestalt associated with the pitch pattern. This pattern is created in that each tonal element of the pitch gestalt is contrasted with the preceding element (within the same gestalt unit) in terms of its pitch, length, and loudness properties. These multidimensional backward contrasts and the concomitant contrast enhancements and top-down expectations create a specific prominence level for each tonal element and, hence, a specific prominence pattern across all elements.

This key mechanism explains effects of synchronization, shape, and their interplay with syllable structure. For example, changes in F0 peak shape cause changes in the duration of tonal elements, and longer tonal elements are perceived as being more prominent than shorter tonal elements. Even more important from the perspective of the current KIM, synchronization is no longer a direct phonological feature in the C-TIP. Rather, synchronization is merely considered *the most efficient and effective strategy for increasing* the relative prominence of a tonal element in the pitch gestalt by exploiting the given high-intensity area of the accented vowel. In terms of the metaphor displayed in figure 8.12b, synchronizing of F0 peaks (or valleys) relative to the segmental string is about moving tonal elements "into the light" of the (high-intensity)

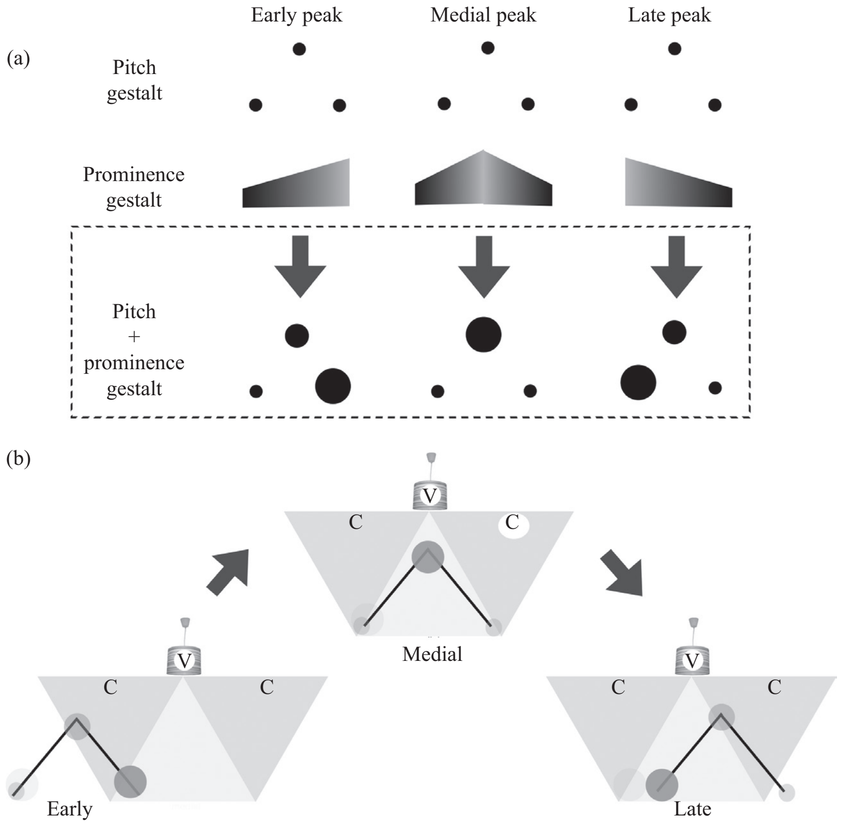


Figure 8.12
(a) Representation of pitch-accent peaks in the contrast theory of intonation perception (Niebuhr 2007b). (b) Implications for the role of synchronization in pitch-accent identification.

vowel section, or out of the light into the "darker" (low-intensity) sections of adjacent consonants.

The metaphor in figure 8.12b also implies, in line with the empirical evidence in section 8.6.1, that the accented-vowel boundaries are important points of reference in intonational phonology. Yet these points of reference are not absolute. There is more leeway for peak/valley synchronization relative to accented-vowel boundaries when the surrounding (preceding or following) consonants are already quite "light," as in the case of approximants. Moreover, pitch accents synchronized closely to the vowel boundaries are identified less clearly when the intensity increases or decreases from or to the surrounding consonants are shallower. Ultimately, speakers can decide not to make use of synchronization at all and instead change the duration and intensity pattern of the segments that coincide with the tonal elements or the shape of the peak. Speakers can also make use of all options at the same time and to different degrees. This individual or context-specific behavior emerges in production data, for example, in the form of a continuum from "F0 shapers" to "F0 aligners" (Niebuhr et al. 2011) or in the alignment changes under different syllable-structure conditions. The assumption of a separate prominence gestalt is also consistent with the creation of pitch-accent specific micro-rhythms (see section 8.6.2) and explains why listeners are able associate rhythmic sequences of monotonous tones with the corresponding pitch-accent categories of early, medial, and late peak.

8.7 Conclusion

The simple intonational concepts of the KIM, as well as of many other intonation models, appear increasingly outdated against our growing understanding of the actual complexity of intonation. Producing and perceiving intonational categories involves a complex interplay of F0 variables and reaches beyond F0 into other prosodic and even segmental parameters. This complexity calls for a shift in intonational modeling from the acoustic to the perceptual domain, in which the multitude of acoustic factors, their variation, and effects may become understandable as simple trade-offs in patterns of perceived pitch and prominence. This idea will shape the future development of the KIM. It is in many respects similar to the intensity-weighted tonal-center-of-gravity approach developed independently by Barnes et al. (2010, 2012, 2020) from an autosegmental-metrical starting point. Implementing and potentially combining these perception-oriented lines of research is undoubtedly worthwhile, but will be challenging long-term issues, next to establishing and refining the phonological categories of intonation, emphasis, and prominence themselves. Linking phonology and modeling will raise the additional question of whether perception-oriented ideas and concepts are applicable in the same way to all paradigms and categories.

References

Ambrazaitis, G. I. 2005. “Between Fall and Fall-Rise: Substance-Function Relations in German Phrase-Final Intonation Contours.” *Phonetica* 62:196–214.

Ambrazaitis, G. I. 2009. *Nuclear Intonation in Swedish: Evidence from Experimental-Phonetic Studies and a Comparison with German* (*Travaux de l’institut de linguistique de Lund* Volume 49). Lund: E-Husets.

Ambrazaitis, G. I., and O. Niebuhr. 2008. “Dip and Hat Pattern: A Phonological Contrast of German?” Paper presented at the Fourth International Conference of Speech Prosody, Campinas, Brazil, May 6–9, 269–272.

Ambrazaitis, G. I., and O. Niebuhr. 2014. “Dip vs. Hat Pattern: Is the Intonation Pattern between Two Pitch Accents a Phonological Feature of German?” Poster presented at Lund Symposium 2014, Lund, Sweden, December 7, 2014.

Arnold, D., P. Wagner, and B. Möbius. 2011. “Evaluating Different Rating Scales for Obtaining Judgments of Syllable Prominence from Naïve Listeners.” Paper presented at Proceedings of the Sixteenth International Congress of Phonetic Sciences, August 17–21, 252–255.

Asu, E. L. 2006. “Rising Intonation in Estonian: An Analysis of Map Task Dialogues and Spontaneous Conversations.” Paper presented at the Fonetiikan Päivät 2006/The Phonetics Symposium 2006, Helsinki, Finland, August 30–31, 1–8.

Auer, P., E. Couper-Kuhlen, and F. Müller, F. 1999. *Language in Time: The Rhythm and Tempo of Spoken Interaction*. New York: Oxford University Press.

Barnes, J., H. Mixdorff, and O. Niebuhr. 2020. “Phonetic Variation in Tone and Intonation Systems.” In *The Oxford Handbook of Language Prosody*, edited by C. Gussenhoven and A. Chen, 125–149. Oxford: Oxford University Press.

Barnes, J., N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel. 2010. "The Effect of Global F0 Contour Shape on the Perception of Tonal Timing Contrasts in American English Intonation." Paper presented at the Fifth International Conference of Speech Prosody, Chicago, IL, USA, May 10–14, 1–4.

Barnes, J., N. A. Veilleux, A. Brugos, and S. Shattuck-Hufnagel. 2012. "Tonal Center of Gravity: A Global Approach to Tonal Implementation in a Level-Based Intonational Phonology." *Laboratory Phonology* 3:337–383.

Barry, W. J. 1981. "Prosodic Functions Revisited Again!" *Phonetica* 38:120–134.

Bloch, B. 1948. "A Set of Postulates for Phonemic Analysis." *Language* 24:3–46.

Boersma, P. 2001. "Praat, a System for Doing Phonetics by Computer." *Glot International* 5:341–345.

Bruce, G. 1977. *Swedish Word Accents in Sentence Perspective*. Lund, Sweden: Gleerup.

Bryant, G. A., and J. Fox Tree 2005. "Is There an Ironic Tone of Voice?" *Language and Speech* 48:257–277.

Bühler, K. 1934. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Jena, Germany: Fischer.

Carlson, R., B. Granström, and S. Hunnicut, S. 1990. "Multilingual Text-to-Speech Development and Applications." In *Advances in Speech, Hearing, and Language Processing,* edited by W. A. Ainsworth, 269–296. London: JAI Press.

de Saussure, F. 1916. *Cours de linguistique général*. Paris: Payot.

Dombrowski, E. 2003. "Semantic Features of Accent Contours: Effects of F0 Peak Position and F0 Time Shape." Paper presented at the Fifteenth International Congress of Phonetic Sciences, Barcelona, Spain, August 3–9, 1217–1220.

Dombrowski, E., and O. Niebuhr. 2005. "Acoustic Patterns and Communicative Functions of Phrase-Final Rises in German: Activating and Restricting Contours." *Phonetica* 62:176–195.

Dombrowski, E., and O. Niebuhr. 2010. "Shaping Phrase-Final Rising Intonation in German." Paper presented at the Fifth International Conference on Speech Prosody, Chicago, IL, USA, May 10–14, 1–4.

Fant, G., and A. Krukenberg. 1989. "Preliminaries to the Study of Swedish Prose Reading and Reading Style." *Speech Transmission Laboratory—Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm* 2:1–83.

Görs, K., and O. Niebuhr. 2012. "Hocus Focus: What the Elicitation Method Tells Us about Types and Exponents of Contrastive Focus." Paper presented at the Sixth International Conference of Speech Prosody, Shanghai, China, May 22–25, 262–265.

Grice, M., and S. Baumann. 2002. "Deutsche Intonation und GToBI." *Linguistische Berichte* 191:267–298.

Grice, M., D. R. Ladd, and A. Arvaniti. 2000. "On the Place of Phrase Accents in Intonational Phonology." *Phonology* 17:143–185.

Gussenhoven, C. 1999. "Discreteness and Gradience in Intonational Contrasts." *Language and Speech* 42:283–305.

Halliday, M. A. K. 1967. *Intonation and Grammar in British English*. The Hague, the Netherlands: Mouton.

Heldner, M. 2003. "On the Reliability of Overall Intensity and Spectral Emphasis as Acoustic Correlates of Focal Accents in Swedish." *Journal of Phonetics* 31:39–62.

Hirschberg, J. 2002. “The Pragmatics of Intonational Meaning.” Paper presented at the First International Conference of Speech Prosody, Aix-en-Provence, France, April 11–13, 65–68.

Hirschberg, J. 2004. “Pragmatics and Intonation.” In *Handbook of Pragmatics*, edited by L. R. Horn and G. Ward, 515–537. Oxford: Basil Blackwell.

House, D. 1990. “Tonal Perception in Speech.” *Travaux de l'institute de linguistique de Lund* 24:7–163.

House, D. 1996. “Differential Perception of Tonal Contours through the Syllable.” Paper presented at the Fourth International Conference of Spoken Language Processing, Philadelphia, PA, USA, October 3–6, 2048–2051.

John, T. 2012. “EMU Speech Database System: Praxisorientierte Weiterentwicklung der Funktionalität, Benutzerfreundlichkeit und Interoperabilität sowie die Aufbereitung des Kiel Corpus als EMU-Sprachdatenbank.” PhD diss., Ludwig-Maximilans-Universität Munich.

Jones, D. 1969. *An Outline of English Phonetics*. 9th ed. Cambridge: Heffer.

Kirsch, A. 2015. “Die Produktion und Perzeption früher, mittlerer und später Gipfel über `mmm.” BA thesis, Kiel University.

Kleber, F. 2006. “Form and Function of Falling Pitch Contours in English.” Paper presented at the Third International Conference of Speech Prosody, Dresden, Germany, May 2–5, 61–64.

Kohler, K. J. 1987. “Categorical Pitch Perception. Paper presented at the Eleventh International Congress of Phonetic Sciences, Tallinn, Estonia, August 1–7, 331–333.

Kohler, K. J. 1991a. “A Model of German Intonation.” *Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung (AIPUK)* 25:295–360.

Kohler, K. J. 1991b. “Prosody in Speech Synthesis: The Interplay between Basic Research and TTS Application.” *Journal of Phonetics* 19:121–138.

Kohler, K. J. 1996. “Labelled Data Bank of Spoken Standard German: The Kiel Corpus of Read/Spontaneous Speech.” Paper presented at the Fourth International Conference on Spoken Language Processing, Philadelphia, PA, USA, October 3–6, 1938–1941.

Kohler, K. J. 1997. “Modelling Prosody in Spontaneous Speech.” In *Computing Prosody*: *Computational Models for Processing Spontaneous Speech*, edited by Y. Sagisaka, N. Campbell, and N. Higuchi, 187–210. New York: Springer.

Kohler, K. J. 2004. “Pragmatic and Attitudinal Meanings of Pitch Patterns in German Syntactically Marked Questions. In *From Traditional Phonology to Modern Speech Processing*: *Festschrift for Professor Wu Zongji's 95th Birthday*, edited by G. Fant, H. Fujisaki, J. Cao, and Y. Xu, 205–214. Beijing: Foreign Language Teaching and Research Press.

Kohler, K. J. 2005. “Timing and Communicative Functions of Pitch Contours.” *Phonetica* 62:88–105.

Kohler, K. J. 2006. “What Is Emphasis and How Is It Coded?” Paper presented at the Third International Conference of Speech Prosody, Dresden, Germany, May 2–5, 748–751.

Kohler, K. J. 2007. “Data Acquisition and Modelling in Speech Communication.” Paper presented at the ParaLing07, Sixteenth International Congress of Phonetic Sciences, Saarbrücken, Germany, August 5–10, 1–6.

Kohler, K. J. 2011. “Communicative Functions Integrate Segments in Prosodies and Prosodies in Segments.” *Phonetica* 68:26–56.

Kügler, F., B. Smolibocki, S. Baumann, B. Braun, M. Grice, S. Jannedy, O. Niebuhr, et al. 2015. "DIMA: Annotation Guidelines for German Intonation." Paper presented at the Seventeenth International Congress of Phonetic Sciences, Glasgow, Scotland, August 10–14.

Ladd, D. R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Landgraf, R. 2014. "Are You Serious? Irony and the Perception of Emphatic Intensification." Paper presented at the Fourth International Symposium on Tonal Aspects of Languages, Nijmegen, The Netherlands, May 13–16, 91–94.

Loevenbruck, H., M. Ben Jannet, M. D'Imperio, M, Spini, and M. Champagne-Lavau. 2013. "Prosodic Cues of Sarcastic Speech in French: Slower, Higher, Wider." Paper presented at the Fourteenth Interspeech Conference, Lyon, France, August 25–29, 1470–1474.

Michalsky, J. 2014. "Scaling of Final Rises in German Questions and Statements." Paper presented at the Seventh International Conference of Speech Prosody, Dublin, Ireland, May 20–23, 978–982.

Mixdorff, H., and O. Niebuhr. 2013. "The Influence of F0 Contour Continuity on Prominence Perception." Paper presented at the Fourteenth Interspeech Conference, Lyon, France, August 25–29, 230–234.

Nash, R., and A. Mulac. 1980. "The Intonation of Verifiability." In *The Melody of Language: Intonation and Prosody*, edited by L. R. Waugh and C. H. van Schooneveld, 219–241. Baltimore: University Park Press.

Newman, D. 2016. "Book Review of Instrumental Studies in Arabic Phonetics." *Phonetica* 73, 79–82.

Niebuhr, O. 2006. "The Role of the Accented-Vowel Onset in the Perception of German Early and Medial Peaks." Paper presented at the Third International Conference of Speech Prosody, Dresden, Germany, May 2–5, 109–112.

Niebuhr, O. 2007a. "Categorical Perception in Intonation: A Matter of Signal Dynamics?" Paper presented at the Eighth Interspeech Conference, Antwerp, Belgium, August 27–31, 109–112.

Niebuhr, O. 2007b. *Perzeption und kognitive Verarbeitung der Sprechmelodie. Theoretische Grundlagen und empirische Untersuchungen.* Berlin: de Gruyter.

Niebuhr, O. 2007c. "The Signalling of German Rising-Falling Intonation Categories: The Interplay of Synchronization, Shape, and Height." *Phonetica* 64:174–193.Niebuhr, O. 2008. "Coding of Intonational Meanings beyond F0: Evidence from Utterance-Final /t/ Aspiration in German." *Journal of the Acoustic Society of America* 142:1252–1263.

Niebuhr, O. 2009. "Intonation Segments and Segmental Intonations. Paper presented at the Tenth Interspeech Conference, Brighton, UK, September 6–10, 2435–2438.

Niebuhr, O. 2011. "Alignment and Pitch-Accent Identification: Implications from F0 Peak and Plateau Contours." *Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung (AIPUK)* 38:77–95.

Niebuhr, O. 2012. "At the Edge of Intonation: The Interplay of Utterance-Final F0 Movements and Voiceless Fricative Sounds." *Phonetica* 69:7–27.

Niebuhr, O. 2015. "Stepped Intonation Contours: A New Field of Complexity." In *Tackling the Complexity of Speech*, edited by O. Niebuhr and R. Skarnitzl, 9–74. Prague: Nakladatelství Epocha.

Niebuhr, O. 2019. "Pitch Accents as Multiparametric Configurations of Prosodic Features—Evidence from Pitch-Accent Specific Micro-rhythms in German." In *A Sound Approach to*

Language Matters: In Honor of Ocke-Schwen Bohn, 321–351. AU Library Scholarly Publishing Services. https://doi.org/10.7146/aul.322.218.

Niebuhr, O., M. D'Imperio, B. Gili Fivela, and F. Cangemi. 2011. "Are There 'Shapers' and 'Aligners'? Individual Differences in Signalling Pitch Accent Category." Paper presented at the Seventeenth International Congress of Phonetic Sciences, Hong Kong, China, August 17–21, 120–123.

Niebuhr, O., P. Jarzabkowska, U. Lorenz, C. Schulz, and F. Sodigov. 2012. "Say It Again, Sam! Phonetic Forms and Functions of Emphatic Reduplication in German." Paper presented at the Sixth International Conference of Speech Prosody, Shanghai, China, May 22–25, 258–261.O'Connor, J. D., and J. F. Arnold. 1970. Intonation of Colloquial English. London: Longman.

Niebuhr, O., and K. J. Kohler. 2004. "Perception and Cognitive Processing of Tonal Alignment in German." Paper presented at the First International Symposium on Tonal Aspects of Languages, Nanjing, China, March 28–31, 155–158.

Niebuhr, O., and H. P. Pfitzinger. 2010. "On Pitch-Accent Identification; the Role of Syllable Duration and Intensity." Paper presented at the Fifth International Conference on Speech Prosody, Chicago, IL, USA, May 10–14, 1–4.

Niebuhr, O., B. Schroeder, and S. Baumann. 2015. "Cues to Perceptual Prominence of Prenuclear Pitch Accents: Evidence from German Spontaneous Speech." In *Proceedings of the International Conference on Prominence in Language*, 1–2.

Niebuhr, O., and M. Zellers. 2012. "Late Pitch Accents in Hat and Dip Intonation Patterns." In *Understanding Prosody: The Role of Context, Function and Communication*, edited by O. Niebuhr, 159–186. Berlin: de Gruyter.

Osgood, C. E., G. J. Suci, and P. H. Tannenbaum. 1957. *The Measurement of Meaning*. Champaign: University of Illinois Press.

Peters, B. 2000. "Individuelle und geschlechtsspezifische Unterschiede in der prosodischen Gestaltung deutscher Lese- und Spontansprache." *Sprache und Kultur, Forum Angewandte Linguistik* 38:153–162.

Peters, B. 2005. "The Database: The Kiel Corpus of Spontaneous Speech." *Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung (AIPUK)* 35a:1–6.

Peters, B. 2012. *Form und Funktion prosodischer Grenzen im Gespräch: Ein phonetischer Beitrag zur Gesprächsforschung*. Saarbrücken, Germany: Südwestdeutscher Verlag für Hochschulschriften.

Peters, B., and K. J. Kohler. 2004. "Trainingsmaterialien zur prosodischen Etikettierung mit dem Kieler Intonationsmodell KIM." www.ipds.uni-kiel.de/kjk/pub_exx/bpkk2004_1/TrainerA4.pdf.

Peters, B., K. J. Kohler, and T. Wesener. 2005. "Melodische Satzakzentmuster in prosodischen Phrasen deutscher Spontansprache: Statistische Verteilung und sprachliche Funktion." *Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung (AIPUK)* 35a:7–54.

Petrone, C., and O. Niebuhr. 2013. "On the Intonation in German Intonation Questions: The Role of the Prenuclear Region." *Language and Speech* 57:108–146.

Rietveld, T., and A. Chen. 2006. "How to Obtain and Process Perceptual Judgements of Intonational Meaning." In Methods in Empirical Prosody Research, edited by S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schliesser, 283–319. Berlin: Mouton de Gruyter.

Ritter, S., and T. B. Roettger. 2014. “Speakers Modulate Noise-Induced Pitch according to Intonational Context.” Paper presented at the Seventh International Conference of Speech Prosody, Dublin, Ireland, May 20–23, 1–5.

Strangert, E. 2003. “Emphasis by Pausing.” Paper presented at the Fifteenth International Congress of Phonetic Sciences, Barcelona, Spain, August 3–9, 2477–2480.

Strömbergsson, S., J. Edlund, and D. House. 2012. “Prosodic Measurements and Question Types in the Spontal Corpus of Swedish Dialogues.” Paper presented at the Thirteenth Interspeech Conference, Portland, OR, USA, September 9–13, 839–842.

Trask, R. L. 2004. *A Dictionary of Phonetics and Phonology*. New York: Routledge.

Trubetzkoy, N. S. 1939. *Grundzüge der Phonologie*. Travaux du Cercle Linguistique de Prague, vol. 7. Prague: Cercle Linguistique de Prague.

von Thun, F. S. 1981. *Miteinander reden: Störungen und Klärungen. Psychologie der zwischenmenschlichen Kommunikation*. Reinbek, Germany: Rowohlt.

Wahlster, W., and J. Engelkamp. 1992. *Scientific Goals and Networks of Work Packages of the Verbmobil Project*. Saarbrücken, Germany: Deutsches Forschungszentrum für Künstliche Intelligenz.

Winchester, S. 2003. *The Meaning of Everything: The Story of the Oxford English Dictionary*. Oxford: Oxford University Press.

Xu, Y. 2004. “The PENTA Model of Speech Melody: Transmitting Multiple Communicative Functions in Parallel.” Paper presented at the International Symposium from Sound to Sense, Boston, MA, USA, June 11–13, 91–96.

Żygis, M., D. Pape, L. M. T. Jesus, and M. Jaskuła. 2014. “Intonation of Statements and Polar Questions in Whispered, Semi-Whispered and Normal Speech Modes.” Paper presented at the Seventh International Conference of Speech Prosody, Dublin, Ireland, May 20–23, 1–5.

9

The Rise and Fall of the British School of Intonation Analysis

Francis Nolan

9.1 Introduction

The British school of intonation analysis blossomed in the first half of the twentieth century and continues to be used today, particularly in language teaching. It builds on earlier insights into the way in which patterns of variation in pitch and other dimensions can be used to convey information additional to the meaning of the words in an utterance. Given the era in which it was developed, the British school of intonation analysis relied mainly on auditory phonetic analysis of the patterns and introspection about their meaning. One of its defining characteristics is the use of dynamic pitch elements such as rise and fall-rise as opposed to pitch levels. Its name reflects the fact that, although contributions have been made by scholars around the world, much of the development of the analyses that characterize the school can be attributed to phoneticians working the United Kingdom.

The name has no formal status, but has been used to contrast the kinds of analysis pursued within this particular general framework with those of other traditions, and the label has been well accepted by those working in intonation. In the mid-twentieth century, the contrast was principally with a framework based on pitch levels that was popularized by work in the United States, whereas nowadays the contrast is more commonly with the autosegmental-metrical (AM) framework that predominates in experimental research into intonation. It should not be thought, however, that the British school is a monolithic, invariant framework in terms of methodology and concepts. This chapter will outline its history and show how those working on intonation edged toward the units and structures that are now associated with the British school, and it will discuss some of the variants that coexist within the broad framework. It will also look explicitly at the relation of the British school to AM models to see what they share, how they differ, and whether there are insights that can be usefully adopted in either direction.

9.2 The History of the British School

The history of the British school is one of incremental progress toward an understanding of the nature of intonation and how to analyze it. No phenomenon, particularly in language, yields to analysis without a struggle, but intonation seems to have put up a particularly spirited fight. Seen from the perspective of our current understanding of intonation, the reason for this arguably lies in its multifaceted nature and its interweaving with other prosodic systems. Several crucial features of intonation that are widely agreed today were not self-evident.

First, intonation functions both within the formal linguistic system and outside it. For instance, intonation can signal the boundaries of structural units such as clauses, and it has the potential to change a statement into a question. Conversely, it can be used to convey paralinguistic meaning outside core language—for instance, information about the speaker's attitude. A failure to appreciate such distinct functions has often led to confusion: a rise on an interrogative is because of the interrogativity, if the analyst's starting point is that intonation is determined by grammar, or it is because of the speaker's wish to sound friendly and unthreatening, if the belief is that intonation's work is attitudinal. Only if it is appreciated that intonation is carrying out both tasks simultaneously can its inner workings be laid bare.

Second, and relatedly, intonation imposes on its primary carrier, pitch, both discrete categories and gradient variation. A rise and a fall are (almost certainly) discretely different, while expansion of pitch range to convey shades of enthusiasm is continuous. To the extent that discrete categories are involved, this suggests that intonation is in some measure phonological. It would be neat and simple if the workload were correspondingly divided—categories for linguistic function, and gradience for paralinguistic work—but in reality, things are more complicated. So, as well as using a pattern ending in a fall to ask an English *wh*-question, as is generally considered the default from the grammatical point of view, we can also end it on a different category, a rise, to increase the level of politeness and make it less face-threatening. Conversely, a declarative can be given the discourse status of a question by expansion of the pitch range of the utterance including the final fall. The mix of discreteness and gradience tended to obscure the matter of whether, and to what extent, intonation shares the kind of phonological structure familiar from segments. The idea that intonation is phonologically structured is one that has only relatively recently gained ascendancy and, even now, uncontroversially so more among linguists than among psychologists and speech engineers.

To prepare the ground for one of the themes that will run through this chapter, let me briefly discuss what makes linguistic analyses "phonological." Ladd (2008, 10) suggests:

> At a minimum, a complete phonological description includes (a) a level of description in which the sounds of an utterance are characterised in terms of a relatively small number of *categorically distinct entities*—phonemes, features, or the like—and (b) a mapping between such a description and a physical description of the utterance in terms of *continuously varying parameters*, such as an acoustic waveform or tracks of the movement of the articulators.

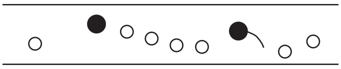
To this one might add an expectation that the categorically discrete entities will have a variable relation both to linguistic function (or "meaning") and to their physical realization and that they can be organized in terms of abstract structures (such as, in today's terms, the *prosodic hierarchy*). While Ladd's definition provides a useful guide to what would be phonological today, Ladd admits that the emphasis on quantitative mapping makes it a characterization of a *laboratory*-phonological approach, a definition that is arguably anachronistic when applied to the earlier phases, at least, of the British school. It has to be remembered that the technology with which we conveniently and rapidly capture, visualize, and quantify the acoustic parameters of intonation postdates the development of the British school. The insights achieved were largely the result of careful listening and introspection.

The same is true of segmental analysis. Classical phoneme theory and Firthian prosodic analysis, for instance, were both eminently phonological, but subject to confirmation largely against impressionistic phonetic judgments rather than instrumental analyses of the acoustic or articulatory events. Ladd concedes that earlier approaches such as the British school and the American structuralist levels framework can be regarded as "proto-phonological," saying "it is important to recognize that the impressionistic descriptions involve phonological categories that could *in principle* be related to instrumentally validated acoustic or articulatory parameters" (12). In this chapter, I will trace the ways in which successive analyses within the British school edged toward a standpoint that can be regarded as fully phonological, judged against the relevant criteria, that is, those appropriate to the contemporaneous methodological state of the art.

A third impediment to insightful analysis of intonation is connected to the earlier lack of clarity about its phonological status and reveals itself in attempts to represent intonation. Intonologists today would generally agree that a fundamental frequency contour does not show intonation, but rather variation in one physical property through which intonation is mapped. The equivalent in impressionistic phonetics would be a wiggly line tracking perceived pitch movement, text meandering up and down the page, or, one might at first suspect, the kind of adaptation of the musical stave frequently used to show relative pitch. All of these succeed in representing one aspect of intonation, the pitch variation, but require considerable ulterior knowledge of other aspects. In particular the first two kinds of representation, and the simplest form of the third, fail to reflect one of the insights that fades in and out of focus in the British school, namely the symbiotic relation between intonation and stress in English. The notion that metrically strong syllables provide anchor points to which intonational landmarks associate is now familiar, but took time to emerge.

The kind of visual representation that evolved during the development of the British school, a generic example of which is shown in (1), reflects a substantial degree of abstract linguistic analysis and understanding of these matters.

(1) There's ↘no need to interˇfere with it



The interlinear graph, or "tadpole diagram," relies on a division of the utterance into syllables and a classification of syllables as stressed (here, the filled dots) or unstressed. Pitch is represented impressionistically in the range defined by the stave, and, traditionally, a judgment is made about which syllable or syllables—most often the nucleus—have perceptible pitch movement. As noted by Cruttenden (1997, xv) this kind of representation is similar to a narrow phonetic segmental transcription. In the text, the symbols—often known as "tonetic stress marks"—show both the division of the intonational phrase into structural elements and the phonological categorization of those elements. Phonological categories such as the fall-rise nucleus on the stressed syllable of *interfere* have their phonetic realization made explicit by the tadpole diagram; the fall-rise would be shown compressed onto a monosyllabic nucleus by a tadpole whose tail bends back upward or extended over a relatively flat "valley floor" in the case where several syllables follow the nuclear syllable. We shall see that, as in other areas of phonetics and phonology, the evolution of representations was gradual and went hand in hand with theoretical progress in understanding intonation.

A general point that needs to be borne in mind about the British school is that the aim of the enterprise was overwhelmingly pedagogical. Those who saw intonation as a vital, but neglected, part of language teaching, mainly in connection with the teaching of English, were discovering the facts of intonation with a view to making them accessible to language teachers and learners. Crystal (1969a), a work that approaches prosodic systems from first principles and aims to give a comprehensive descriptive account of both linguistic and paralinguistic aspects, is a notable exception to this goal-oriented approach, but for most authors there was a tension between the intellectual satisfaction of progressively uncovering complexity in the prosody of language and the need to systematize, simplify, and codify what was emerging for the practical purpose of language teaching.

In next subsections I will give a selective summary of the history of the British tradition, focusing on its zenith in the twentieth century, but first considering its roots in earlier work. Readers who want a comprehensive and wide-ranging account should consult Crystal (1969a); here it is only possible to focus on notable milestones in the development of the framework. Cruttenden (1997) provides a wide-ranging introduction to intonational description and theory grounded mainly within the British framework, and Cruttenden (1990) gives specifically the genesis of the concept of the intonational nucleus. Where appropriate I will refer back to the matters that have already been raised: the phonological status of intonation, its linguistic and paralinguistic roles, the interplay of categories and gradience, and the kinds of representation used by different authors. On the whole, the bias will be toward the identification and representation of intonational form rather than the attempts to pin down the function of the forms identified.

9.2.1 Before the Twentieth Century

The earliest works discussing the intonation of English, according to Crystal (1969a, 20–22), were those of John Hart in the sixteenth century. In *The Opening of the Unreasonable Writing of Our Inglish Toung* (1551), Hart is concerned with the nature of stress and phrasing, offering diacritics to improve the relationship of writing and speaking. His *Orthographie* (1569) discusses intonation in the context of punctuation marks, identifying the common falling intonation pattern of commands, exclamations, and *wh*-questions, and giving a detailed discussion of phrasing. Butler ([1633] 1910) has a brief discussion of "tone" in connection with "points" (punctuation marks), *tone* being "the natural and ordinary tune of the voice: which is to rise, or fall, as the Primary points shall require: and therefore it denominateth the voice, High or low" (54).[1] Butler's observations are few but show insight, for instance: "Erotesis [the question mark], if it be pure, raiseth the common Tone or tenour of the voice in the last word; unless Emphasis draw it: but if it begin with a word interrogative; as, [who, what, how, where, why, &c;] it falleth as a Period, and raiseth the tone in the Interrogative" (61). I am not sure what is meant by "raiseth the tone in the Interrogative"—maybe he is thinking of a rise at the end of echo questions—but Butler has clearly recognized that not all sentences written with ? rise at the end. Having already discriminated (60) three uses of the question mark[2]—"asking," "urging or instance in reprehension," and "earnest avowing the contrary"—Butler suggests that in the latter two (61) erotesis "straineth the voice throughout the whole Interrogation," perhaps a reference to increased overall pitch range and loudness.

Steele ([1775] 1969) argues the case for English actually having intonation; this is in response to Lord Monboddo (James Burnett), whose view was that English "is altogether unmusical, unless we are pleased to call a drum a musical instrument—For it has no melody, that is tones, differing in acuteness and gravity upon different syllables"

(Burnett 1784, 113).[3] Steele's methodology and notation are impressive, albeit he perhaps fails to see some of the phonological wood for the phonetic trees. He used a base viol to capture pitch glides as follows:

> Let a bass viol have paper pasted on the whole length of the finger-board near the 4th string, marking all the chromatico-diachronic stops or frets. . . . Then, while the player draws the bow over the 4th string, let him try, by sliding his left hand on the same string up or down the finger-board, to imitate the rapid turns or flections of the voice in common speech, and he will soon find, that they will have either their beginnings or endings, for the most part, in the interval between the fretts; which intermediate stops, we may call quarter tones: for it will be accurate enough for our purpose, to call every degree of tone a quarter, that does not coincide with any tone or semitone of the chromatico-diatonic. (15–16)

He says in summary that speech is not monotonous like a drum; that rises and falls ("changes of the voice from *acute* to *grave*, and *vice versâ*") do not correspond to musical intervals and need quarter tones for their description; that glides occur not just on words but syllables and monosyllables; and that such glides are in the order of a (musical) fifth.

Steele distinguishes five properties of what we would call prosody (viii): *accent*, *emphasis*, *quantity*, *pause*, and *force*. His notation is exemplified in figure 9.1; this example (51) is in a "bombastic" rhetorical style. Each syllable is represented by a stylized musical note (in this example, minims, as for "now," "need," and "calls," crotchets—the majority of syllables—and quavers, as for the first three syllables of "deliberation," here forming a triplet), the pitch glide (the *accent*) being represented by the sloping line within the stave. The importance of pauses is understood, and it is tempting to attribute to Steele a recognition of final lengthening (for instance, in this example, on "now" and "need"). *Emphasis* contributes to what we would recognize as speech rhythm, with syllables being "heavy" Δ, light ∴, or lightest .., and organized into *cadences*, maybe rather like feet. He distinguishes *emphasis* in this sense from *loudness*, in the latter the "application of the *forte* and *piano* being *ad libitum*," and therefore in contemporary terms not phonologically determined but chosen stylistically. *Crescendo* (and its reverse, *smorzando*, not present in this sample) are shown iconically by wiggly lines, with ' indicating *piano* on a syllable and ' *forte*. Although this notation is both original and unadopted by others, in its principles it clearly prefigures the kind of intonational representation that would become familiar in the twentieth century—and embraces a range of prosodic features that is scarcely matched until Crystal (1969a).

A more immediate precursor to the British school, a century after Steele, is that preeminent scholar of English philology and phonetics, Henry Sweet, whose *Handbook of Phonetics* ([1877] 2013) and *Primer of Phonetics* (1892) include some observations about intonation, despite his declaring, "The whole subject of intonation, especially, requires to be thoroughly investigated by a thoroughly competent observer, which I am very far from being, my natural aptitude and my training being equally defective" ([1877] 2013, x).

Sweet is acutely aware that speech does not segment itself in a way corresponding to its lexical structure, noting "the only division actually made in language is that into 'breath groups'" (86). He suggests that for practical purposes it is sufficient to recognize three degrees of stress, while in reality "the variations of stress are infinite" (92), giving example (2), which almost seems to prefigure the output of the stress cycle of Chomsky and Halle (1968):

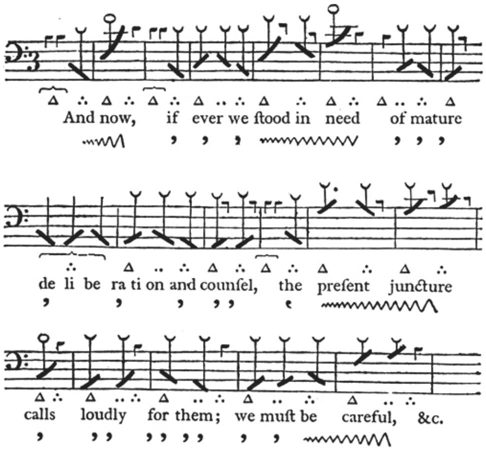

And now, if ever we ſtood in need of mature
de li be ra ti on and counſel, the preſent junĉture
calls loudly for them; we muſt be careful, &c.

Figure 9.1
An illustration of the prosodic representation used by Steele ([1775] 1969).

(2) 2 3 7 5 1 6 4
im – pe – ne – tra – bi – li – ty

As far as intonation is concerned, Sweet recognizes three primary tones, *level*—, *rising* /, and *falling* \, together with *compound rising* and *compound falling* (i.e., fall-rise and rise-fall), and even a rise-fall-rise as a "somewhat more emphatic" version of the fall-rise ([1877] 2013, 94). He is also very aware of the independence of pitch range: "All these tones can be varied indefinitely according to the interval through which they pass. As a general rule, the greater the interval, the more marked the character of the tone" (94). Virtually the same wording is used in Sweet (1892, 66), with the proviso that "for ordinary purposes," it is enough to distinguish between high and low versions of the rise and fall, a solution that is adopted in some later influential works (notably O'Connor and Arnold 1961). In transcription, tone marks "would take the place of the ordinary marks of punctuation at the end of each breath-group, thus serving the double purpose of indicating breath-division and intonation" (Sweet [1877] 2013, 107), as in (3), which is represented (115) in "Broad Romic" and glossed here orthographically for convenience. Spaces precede primary stressed syllables, the raised dot indicates a secondary stress, the low dot a pause (here between interlocutors, I presume) and the colon indicates an initial secondary stress.

(3) :dyuwəb jekt(t)ətə bækos·mouk / . notə taol \
‘D’you object to tobacco smoke?’ ‘Not at all.’

Sweet's inventory of tones would not look out of place in a mid-twentieth-century British analysis, and there is an implicit recognition of the foot. However, in its lack of an explicit association, as we would call it today, between the tone and a metrically strong syllable, the notation leaves phonological detail to the knowledge of the speaker in a way that later notations do not. The "prenuclear" intonation pattern, on the material before the tone, is also left unspecified except insofar as stress is indicated.

We shall see that a significant insight of the British school is the structural analysis of the intonational phrase,[4] as it would be known today, into subelements where contrastive choices can be made. It is apparent that a degree of structural analysis is implicit in Sweet's focus on what would later be termed the *nucleus*, but the structure is left to his successors to develop, as is set out in more detail by Cruttenden (1990).

9.2.2 The Early Twentieth Century

9.2.2.1 Jones Daniel Jones, the dominant figure in British phonetics in the first half of the twentieth century, published *Intonation Curves* in 1909. This was methodologically innovative, in that he describes a method using gramophone recordings and lifting the stylus because "the ear will retain the impression of the sound heard at the instant when the needle is lifted" (v). The pitch of this impression he would then compare with a tuning fork and mark on a musical stave, in a practice reminiscent of Steele. By comparison with Steele, however, the analysis is less sophisticated, because while stress, vowel length, and segmental quality are recorded in a segmental phonetic transcription, the intonational representation consists merely of a continuous line (dotted for voiceless episodes) meandering up and down a stave, with bar lines separating syllables (see figure 9.2). The result is therefore the auditory equivalent of a fundamental frequency contour, and so a raw pitch curve with no attempt either to make phonological abstractions or group patterns according to function.

A similar line representation is used in early editions of Jones's *Outline of English Phonetics* (e.g., 1918), though the curve is on a stave that has lost all but its top and bottom line, and musical intervals are indicated by a separate musical stave with a lower and a higher note indicating the excursion of the pitch change. Here, there is more analytic discussion, linking examples to both grammatical and attitudinal functions. The numerous examples are organized under rules, such as "IIIrd rule. Direct questions capable of being answered by 'yes' or 'no' generally have a rising intonation at the end" (section 709), and lists of exceptions; but the organization is by function rather than by systematic analysis of the pitch patterns, and as a consequence there is a lack of generalization. After the first two editions in the long life span of the *Outline*, Jones moved to a much simpler representation more in keeping with later practice in the British school. This used interlinear graphs, unusually with one mid-pitch line in addition to the top and bottom lines; larger dots indicated stressed syllables and smaller dots unstressed syllables. A functional analysis in terms of two basic tunes is also adopted, which he explicitly credits to Armstrong and Ward.

9.2.2.2 Armstrong and Ward The first half of the twentieth century saw models being developed that continued to focus on the crucial tonal event toward the end of the intonational phrase, while others began to extend their scope to contrasts elsewhere and move toward a structural analysis of the whole intonational phrase. An example of the first type of model is Armstrong and Ward (1926), a work that is explicitly

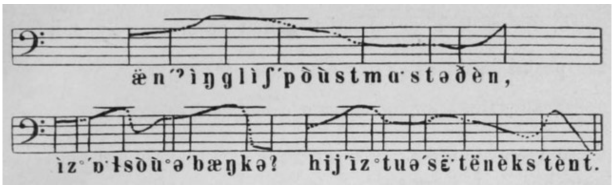
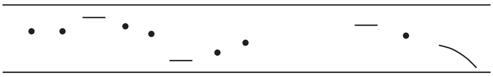

ǣn'ˀìŋglìʃ'pòùstmɑˑstəðèn,
ìz˚'ɒˑɫsòù˚ə'bæŋkə? hij'ìz˚tuə'sɛ̈ˑtënèks'tènt.

Figure 9.2
An illustration of the intonational representation used by Jones (1909, 21).

pedagogical and simplifying. They observe that “English intonation can be reduced to two tunes, with variations of these due to special circumstances” (4). The tunes apply to “intonation groups,” which correspond to the “sense-groups [hyphenation thus]” that make up connected speech. In tune I, “the stressed syllables form a descending scale. Within the last stressed syllable, the pitch of the voice falls to a low level.” Tune II follows “the outline of [tune I] until the last stressed syllable is reached. This is on a low note, and any syllables that follow, rise from this point. . . . If the last stressed syllable is final, the rise, which is an essential of this tune, occurs within the stressed syllable itself” (20). In the notation used by Armstrong and Ward, the example (3) quoted from Sweet might have been analyzed as shown in (4) (for simplicity I have replaced the segmental phonetic transcription that is used throughout the work by orthography):

(4) D’you obˈject to toˈbac co smoke? ˈNot at ˈall.



In the text, sentence stress is marked by the prime ˈ symbol, and in the interlinear representation stressed syllables are shown as lines rather than dots, with any curvature of the line indicating a perceived pitch glide on a syllable. Secondary stress is not marked (Armstrong and Ward 1926, 57; see, e.g., [ˈreɪlweɪ steɪʃn] [79]), though may reveal itself through pitch movement on a final secondary-stressed syllable. In (4), the first part of the dyad carries tune II, the rise being realized as steps over the primary-stressed syllable (*-bac-*) and the following two syllables; and the response carries tune I, with the fall realized as a glide over the final monosyllable. Although it is not discussed in such terms, this insight into the equivalence of patterns over different syllabic material reflects the phonological nature of the analysis, namely that an abstract element in the system (e.g., a rise) can have variable realization depending on the phonetic context.

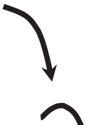
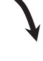
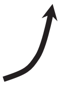
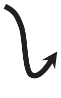
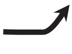
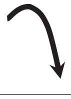
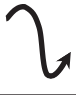
This basic analysis is elaborated by a discussion of sentences with special emphasis. This includes the use of expanded pitch range, more complex pitch patterns including fall-rise, rise-fall, and rise-level, and (as we would understand it today) earlier placement of the nucleus with postnuclear deaccenting. In fact, most of the patterns familiar from later work are exemplified. However, the basic categorization into tune I and tune II is maintained throughout, without formal recognition of an inventory of nuclei or an internal structure to the intonational phrase, suggesting priority is being given to a functionally driven analysis of the kind correlating with the final pitch movement. Cruttenden notes of his own inventory of seven nuclear tones, and by implication those of others, that the inventories are generally reduced (or perhaps better abstracted away from) by collapsing a number of nuclear tones under one heading: “Typically, we end up with either a simple distinction between falls and rises, or else we are presented with a threefold distinction between falls, rises, and fall-rises” (1997, 106). In their focus, formally, on a twofold intonational distinction based on the final pitch movement, Armstrong and Ward (1926) are presenting a more broad-brush characterization of English intonational forms than is found later in mid-twentieth-century analyses—albeit one that prefigures the use of boundary tones in current intonation analysis.

9.2.2.3 Palmer It is with Palmer (1922) that we see a major step toward a conception of intonation as a phenomenon with greater phonological structure. As others had done implicitly, he recognizes explicitly and names the *nucleus*, defining it as “the stressed syllable of the most prominent word in the Tone-Group" (7), and he also innovates the term *head* for "any syllable or syllables preceding the nucleus in the same Tone-Group."[5] By doing so, Palmer gives the intonational phrase a structure and is able (i) to talk about significant choices—the phonological primes—available at both those places in structure; (ii) to suggest restrictions on co-occurrence between those phonological elements (for instance, "it is doubtful whether the low-rising nucleus (Tone-Group 4) is ever preceded by an inferior head" [28]); and, as we shall see, (iii) to recognize phonologically insignificant variation in the realization of the phonological primes.

He also refers to any syllables in the tone group occurring after the nuclear syllable as the "tail" (10–11), but he makes it clear that the pitch of the tail is determined by the choice of nucleus. The lack of phonological choice in the tail, which most subsequent authors agree on, is an inconsistency if the point of the structural analysis of the intonational phrase is to identify positions at which significant choices can be made. It is, however, useful to have a term to refer to that position because, pedagogically, there is much to learn about the realization of nuclear tones depending on the type of material, if any, in the tail.

Palmer recognizes four "characteristic Nucleus Tones": *falling*, *high-rising*, *falling-rising*, and *low-rising*. Each defines a tone-group type, showing that Palmer adheres to the notion familiar throughout the British tradition that it is the nucleus that gives the intonational phrase its main meaning. It is not made explicit what the criteria are for "high-" and "low-," but it seems from the graphical representation of patterns that it is more to do with the endpoint of the rise rather than the start. It is clear that he regards the two kinds of rise as phonologically distinct, rather than merely being variants of realization, because he comments separately that "the *Range* of a falling or rising tone varies according to the degree of animation of the speech". Perhaps surprisingly Palmer does not distinguish in his inventory between two heights of fall, as some subsequent authors do, yet he employs two distinct symbols (see table 9.1) of different span whose use, to judge from the examples given, is determined by the pitch relationship between the head and the nucleus. He also allows for "Intensification" of the falling nucleus where "the actual fall is preceded by a slight rise of pitch," while it seems that the falling-rising tone in the inventory is "almost invariably intensified" (8) and has a slight leading rise, but occasionally may merely fall and rise. Table 9.1 approximately reproduces Palmer's symbols for annotating texts.

Table 9.1
Palmer's (1922) nucleus tone inventory

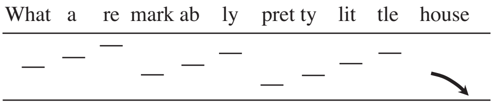
Tone group	1	2	3	4
Nucleus	Falling	High-rising	Falling-rising	Low-rising
Nonintensified	 			
Intensified				

Note: The nonintensified falling tone is the default, whereas the intensified falling-rising tone is the one normally encountered, according to Palmer.

Palmer recognizes four kinds of head: *inferior*, *superior*, *scandent*, and *heterogeneous* (17). The definition of the first three involves the relation of the pitch of the head relative to the nucleus; the pitch of syllables in the inferior head are never higher than the start of the nuclear tone, while that of the superior head is higher. The pitch of the scandent head climbs from a mid pitch to the highest point of the tone group. These relational definitions allow of variation in absolute pitch, and the notation for the head is more phonetic than phonological. For instance, —ai ⤵siː and _ai kæn⤴ both have inferior heads, the horizontal line indicating the pitch of the head relative to the nucleus, while —nekst ⤴wiːk and ¯nekst ⤴wiːk both have superior heads, the difference being attributed to "the degree of intensity with which the tone group is uttered" (17–18).

In longer heads, there is even more scope for realizational variation: for instance, the superior head involves a general decline in pitch over a sequence of syllables, one that may describe an overall slope, or a stepping pattern with downstep (as it might be described today) on each stressed syllable with unstressed syllables retaining the pitch of the preceding stressed syllable. Realizational variation is particularly striking in the scandent head. Over longer material, there is gradual rise from a mid pitch to the highest pitch of the tone group. However, in longer scandent heads the rise can be *broken* one or more times (46), that is, the pitch is reset to a lower pitch than the initial pitch and another rise begun, giving a kind of sawtooth pattern, as in (5)[6]:

(5)



Heterogeneous heads involve the combination of any of the other three types. Palmer admits that he has little to say about the purpose of such combinations, but suggests "the varying head-curves correspond more or less to the various degrees of prominence expressed by each significative element in the head." His intriguing suggestion is that "many of these heads may perhaps represent embryonic nuclei" (69). This could even be interpreted retrospectively to be an intuition that a lower level of phrasing might be involved. At the very least there is a hint that not only is the tone group decomposible, but also the head. The former point, but not the latter, finds favor generally within later British school approaches.

Having dealt with intonational form, Palmer goes on to deal—albeit somewhat apologetically given, as he says, the (prevailing) state of knowledge—with the "semantic functions" of the four tone groups and the effect of combinations of nucleus and head. He is concerned with grammatical and discourse determinants, such as "general" (yes/no) and "special" (*wh*-) questions, but also offers comments on attitudinal effects, for instance of the combination of tone group 1 with heads: "The significative difference between the Superior and Scandent Heads is difficult to define with precision. As compared with a Superior Head, the Scandent Head generally expresses more animation" (76–77). He is even keen to seek explanations and cites approvingly Coleman (1914) who suggested that general questions have rises (tone group 2) because they are alternative questions in which the second alternative has been suppressed. Palmer is also alert to the balance of prominence within the intonational phrase, pointing out for instance that with the combination of inferior head and falling nucleus, the "prominence is almost always confined to the nucleus-word," whereas "in a Tone-Group containing a superior head, the prominence is distributed over the head and the nucleus-word, the nucleus (as by definition) having the maximum of prominence" (73, 75). This, whether inadvertently or wittingly, anticipates one of the problems bedeviling the British school, namely the failure of the nucleus or tonic to be self-evidently the most prominent syllable in the intonational phrase (cf. Brown, Currie, and Kenworthy 1980).[7]

I have dealt with Palmer at length because he seems best among the early authors to lay the foundation for a fully phonological approach to intonation modeling, one that, I am claiming, is characteristic of the British school. As noted, he introduces structure to the intonational phrase and recognizes the importance of whether variants found in those positions of structure are "significant" (this is not to say that he has a watertight test for this crucial difference, any more than intonologists today have). He also briefly considers co-occurrence restrictions between choices in the head and at the nucleus—an intonational phonotactics, as it were. In terms of the functions of intonation, he is responsive to attitudinal meanings as well as to the association of intonation with grammatical and discourse functions and, one might venture to claim retrospectively, to information structure when he talks about the balance of prominence in different combinations of head and nucleus. At the same time, he fails to be explicit about one tenet that characterizes the "standard" British model, namely that significant pitch movements are analyzed as beginning on stressed syllables. His nucleus indeed has to be a stressed syllable, but his head symbols are be placed at the start of the sentence, regardless of whether those syllables are stressed or unstressed.[8]

9.2.3 The Heyday of the British School

9.2.3.1 Kingdon If there is a period that could be identified as the one in which the refinement and influence of the British school was at its height, it would be the third quarter of the twentieth century. The postwar period saw the appearance of Kingdon's *The Groundwork of English Intonation* (1958), which the author says "is in part based on one of the systems used by Harold E. Palmer, but . . . contains a considerable number of additions and modifications that I have found it advisable to make" (xviii). The system referred to is the one in Palmer (1922) just discussed, which recognizes the independence of the head, and not that of Palmer (1933), which moves toward a description in terms of whole tunes—the latter being a retrograde step within the overall evolutionary trend described here. Kingdon is critical of the whole-tune approach, attributing the inferior progress of English learners' intonation by comparison with their segmental pronunciation to their being made simply to repeat sentences "after a superficial description of two contrasting tunes. . . . It seemed to me that a book in which the essential elements of intonation were presented and described separately, and then assembled into tunes, might enable a considerable advance to be made in teaching methods, just as the use of phonetic analysis has greatly improved the teaching of pronunciation" (xv).

Kingdon's analysis recognizes five *kinetic tones*. These are most commonly associated with nuclear syllables and in that position will be realized in various ways over the tail following the nuclear syllable; though they can also occur in prenuclear stretches. The *nucleus* is the tone associated with the last fully stressed syllable of a tune, a definition that sidesteps the difficulty encountered by formulations that rely on the nucleus being the most prominent element of the intonational phrase. His definition of *stress*, "force employed in uttering a syllable," is less felicitous, as is his view that stress varies within a syllable, reflected in his idiosyncratic use of a "wedge" in the interlinear graphs to represent stressed syllables to "suggest the relative degree of stress at different points" (4).

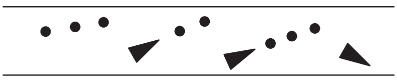
The kinetic tones are *rising*, *falling*, *falling-rising*, *rising-falling*, and *rising-falling-rising*, the first three being considered sufficient for a "straightforward, simple type of English intonation." Slightly confusingly the five kinetic tones are numbered I to V in that order, with the result that a tune ending in a fall is tune II and in a rise, tune I, the reverse of previous tune-based accounts. The reason for this is to be able to group together the three odd-numbered tones ending in a rise and the two even-numbered tones ending a fall, in recognition that "it is the final (not the initial) direction that a tone takes that gives it its principal character" (6), an insight which recurs, and which, as noted in connection with Armstrong and Ward's analysis, chimes with an analysis in terms of high and low boundary tones.

Thus far, the system appears to represent a phonological inventory, with realizational variation arising from the distribution of the pitch movement over different numbers of syllables. However, the picture is somewhat clouded by two cross-cutting dimensions: each of the kinetic tones can occur *low* as well as in its default *high* version, and *emphatic* as well as *normal*. At this point it seems we might be being presented with a twenty-tone inventory, but it is fairly clear that these orthogonal dimensions are largely paralinguistic in function, for instance, "the Low Emphatic tones are used mostly in utterances having an impatient or dramatic nature" (xxxi).

In terms of the structure, Kingdon adopts the term *head* for "the first full stress of a group" (22)—as long as (by implication) the group does not consist merely of the nucleus—and he refers to the remaining syllables, stressed or unstressed, before the nucleus as the *body*. In the case where a prenuclear pattern reduplicates itself, or (in very animated speech) several kinetic tones precede the nucleus, this seems to have the disadvantage of divorcing any unstressed syllables after the "first full stress" and grouping them, in the body, with the following stress. But the basic insight that there are contrasts to be made in the prenuclear part of the intonational phrase independent of those of the nucleus is fully worked out. Kingdon also introduces the term *prehead* for any unstressed or partially stressed syllables before the head. The *normal* prehead is produced level and "a trifle above the bottom of the normal voice range" (12); and alternative choices are *high* or *low level* preheads (31) and *ascending* or *descending* preheads (51), which are also possible *high* or *low*. As with the nuclear tones, it seems that here the descriptive framework is capturing potentially phonological contrasts (most authors would agree that "good-bye" with a high unstressed "good" and a low rise on "bye" is discretely different in its rather friendly affect from the same utterance with a low prehead, which sounds grudging), but beyond that, it is likely in retrospect that stylistic realizational variation is being modeled within the categorical inventory. With Kingdon's work, we have arrived at the maximal structure of the intonational phrase within the British tradition. Subsequent work largely agrees on the following structural positions where choices can be made, listed in order of occurrence: *prehead*, *head* (subsuming "body," a term that fades out), and *nucleus(+tail)*.

The example in (6), which I have constructed for comparison with (5) in the discussion of Palmer's (1922) representations, shows several features of Kingdon's analysis. Importantly, the terms in his inventory can for practical purposes be indicated in texts by *tonetic stress-marks* "indicating stress by [their] presence and tone by [their] form" (xxiii), the latter iconically. The first three syllables in the example, none of them fully stressed, constitute a high ascending prehead, clarifying the relation between stress and the head. Kinetic tones in practice are free to occur in nuclear position and prenuclearly, which could be argued to weaken the case for a distinction between nucleus and head (head + body in Kingdon's terms), and to speak in favor of the "flat" structure of the intonational phrase adopted in AM following Pierrehumbert (1980). In contrast, the standard view in the British school would be that even if the inventory in both positions is the same, realizational differences may justify a distinction for pedagogical reasons; in particular, a falling head lacks the sharpness of fall associated with a falling nucleus.

(6) ˊWhat a re ˏmarkably ˏpretty little ˎhouse



This summary can only give a hint of the complexity of Kingdon's formal analysis, which is related by examples to functions of different types. As I have implied, the analysis is perhaps driven more by detectable differences in intonational form than by a strongly phonological impulse to identify all and only the contrastive patterns, but it nonetheless gives an impressive repertory of forms whose distinctive function could, in principle, be tested. His analysis is clearly influential in later works that seek to constrain the combinatory possibilities with the goal of providing more tractable pedagogical accounts of English intonation.

9.2.3.2 O'Connor and Arnold O'Connor and Arnold (1961) has perhaps been the most widely used description of intonation aimed at the foreign learner of British (received pronunciation or RP) English. The authors acknowledge (vii) that they are building on the work of Palmer and Kingdon, though developing it in certain ways. They adopt a structural division of the intonational phrase, or *sense group* as they refer to it, into prehead, head, nucleus, and *tail*. Kingdon's distinction between head and body isn't mentioned, and, as is commonly found in such analyses, it turns out that no independent contrasts are available in the tail. The nuclear tone is realized over the tail in ways that are determined by the nuclear type, the number of syllables available, and whether there are stressed syllables in the tail. This contextual variation of the abstract phonological elements is exemplified in detail.

Six contrasting nuclei are recognized: *low fall*, *high fall*, *rise-fall*, *low rise*, *high rise*, and *fall-rise*. Choice in the head is limited to three types: *low*, *stepping*, and *sliding*. The first syllable of the stepping head is high, and if the head is long enough, it manifests a stepping pattern over two or more accented syllables (like a sequence of downstepped H* pitch accents in AM descriptions); equivalently the sliding head may be realized over longer material as a sawtooth pattern. The low head has two contextual variants, one remaining low throughout and the other occurring only before a high fall, interpolating between low and the high start of the nucleus. Preheads may be high or low. Examples in the tutorial sections are represented using interlinear graphs, large dots for stressed syllables, small for unstressed, and tadpole tails on dots for syllables with perceptible pitch movement. Tonetic stress marks, similar to those of Kingdon, are used for the examples where (in considerable detail) the meaning of the different patterns is discussed and in the extensive chapter with drills for the student.

A particular combination of the four structural elements (with minimally just the nuclear slot, which is obligatory, filled) is referred to as a *tune*. Given the optionality of elements other than the nucleus, and given the dominance of the nucleus in determining the "flavor" of an intonational phrase, the authors group together tunes "all conveying the same attitude on the part of the speaker" (29) into *tone groups*. This is the recognition, in effect, that the intonation pattern is constrained by the amount of phonetic material available. The phonological credentials of the model are supported by the provision (vi) of what is in effect a table of phonotactically permissible combinations in the structural positions, reproduced as figure 9.3. For instance, the sliding head can only occur before a fall-rise nucleus. It is not clear, however, to what extent this statement of co-occurrence restrictions is based on research, on introspection, or is determined by the need to provide a limited set of possibilities to keep the learner's task within bounds.

In their preface, O'Connor and Arnold note:

> In the past, much has been written about English intonation in terms of sentence structure. If, in Chapter II, we appear to have emphasised the relation between intonation and speaker's attitude, it is merely that we have sought to redress the balance and to show that sentence structure and speaker's attitude both play a very important part in determining intonation pattern. (viii)

To that end, they take the four discourse functions *statement*, *question*, *command*, and *interjection*; discuss the effect of each tone group; and exemplify using the constituent tunes of the tone group. For instance, "by using Tone Group 7 with special [i.e.,

	Pre-head		**Head**				**Nuclear tone**						**Tail**
	Low	High	Low	Stepping	Sliding	High fall(s)	Low fall	High fall	Rise-fall	Low rise	High rise	Fall-rise	
Tone Group 1	•		•				✓						•
Tone Group 2	•			✓			✓						•
		✓					✓						•
Tone Group 3	•		•					✓					•
Tone Group 4	•			✓				✓					•
	•					✓		✓					•
		✓						✓					•
Tone Group 5	•			•					✓				•
Tone Group 6	•		•							✓			•
Tone Group 7	•			✓						✓			•
		✓								✓			•
Tone Group 8	•			•							✓		•
Tone Group 9	•				•							✓	•
	•					•						✓	•
Tone Group 10	•		•					✓		✓			•
	•			•				✓		✓			•

Figure 9.3
Co-occurrence restrictions between different choices at places in the structure of the intonational phrase (after O'Connor and Arnold 1961, vi). A tick (check mark) indicates an essential element of the tone group, and dots indicate optional elements.

wh-] questions the speaker seeks to establish a bond with the listener, to show interest not only in receiving the information asked for but also in the listener himself" (54).

The second edition (1973) of *Intonation of Colloquial English* stays true to the broad descriptive framework of the first, but the details are quite substantially changed. Heads are now subcategorized as unemphatic and emphatic; the kind of reduplicating patterns seen on longer material for the *stepping* and *sliding* heads are now regarded as emphatic versions of a *high* head and a *falling* head, and they are joined by a reduplicated version (the *climbing head*) of a new *rising head*, which is no longer regarded as a contextual variant of the low head. It seems that a structurally anomalous description[9] included in the first edition (1961, 26) of a compound tune consisting of a high fall or fall-rise nucleus preceded by one or more falling nuclei (e.g., "You know ˋperfectly ˋwell I ˋnever alˋlow it") can be surreptitiously retired,[10] its work presumably being taken over by the sliding head. However, the compound tune consisting of a high fall and a low rise, which falls and then starts rising on a secondary stress in the tail, is retained (e.g., "My ˋmother lives in ˏSheffield"). This fall + rise pattern poses a perennial problem for the British school because of its structural exceptionalism, but O'Connor and Arnold (1973) retain it, despite admitting that it is often realizationally neutralized with a fall-rise followed by a tail with a secondary stress, because "the two tunes are very different in their meanings" (28).[11] A new *level* nucleus is also introduced. Because of the difficulty of remembering tone groups designated by number (1–10), these families of functionally equivalent tunes, now somewhat revised, are given iconic names such as the *switchback* (the fall-rise nucleus optionally preceded by the falling head and/or a low prehead) and the *long jump* (a high fall, with optional high head and/or low prehead). Schematic interlinear graphs with lines are used to represent the tunes abstractly, in addition to tadpole diagrams for concrete realizations.

The second edition highlights the difficulty all analysts encounter negotiating the terminological quagmire of stress and accent. In O'Connor and Arnold's first edition (1961), a coherent position is almost achieved—but not quite. Words are made prominent by virtue of an accent, an accent resulting when pitch features are combined with stress[12] on a syllable—except, that is, for the low head, where "words are singled out for attention—or accented—by means of stress alone, with no help from pitch features" (17). Admittedly, the initial discussion of the pitch features is in the context of the nucleus, but the implied definition of accentuation logically extends to all stressed syllables associated with a pitch change (but not to those *without* a pitch change, inside a head or a tail, which are just "stressed"), unless apparently they are in a low head. The water is further muddied in the second edition, where any stressed syllable in the head is defined as accented, and pitch obtrusion is regarded as emphasis (as in the stepping, sliding, and climbing heads). My own view on the best line to take pedagogically is that stress in English is a lexical property, a position in a word susceptible of bearing prosodic prominence when it is placed in an utterance. Minimally such prominence where present is rhythmic, realized by greater duration and amplitude on the stressed syllable and potentially some compensatory shortening of adjacent syllables. Such syllables are *rhythmically accented*; syllables made prominent solely in this way most commonly occur in the tail and are indeed normally the only prominent syllables that can occur in the tail. Further prominence may be, and usually is, lent to stressed syllables up to and including the nucleus by pitch, either a kinetic "tone" or pitch obtrusion whereby the syllable steps out of the prevailing trend (see Bolinger 1958). Such syllables are pitch accented, a status that entails that they are also rhythmically accented.

9.2.3.3 Halliday Halliday (e.g., 1967) approaches English intonation from the starting point of his *systemic grammar* and is careful not to claim to be providing an exhaustive account of the phonological contrasts of intonation:

> The system used here reflects the general principle that only those distinctions that are shown in the grammatical description to be meaningful are represented in the phonological analysis . . . [however]. . . . In treating intonation grammatically I am not suggesting that the view of it as carrying emotive meanings is to be rejected. (47)

The analysis is explicitly phonological: "We . . . should set up a single independent phonological system irrespective of the very many different roles that are played by [different selections of] its terms in English grammar." For instance, tone 1 will be the exponent of terms in different grammatical systems, and (rather confusingly put, but interpretable as a statement of an allotonic principle) "is phonetically identical—that is, has the same range of phonetic variety—in all its uses" (11). To discuss the intonational exponence of terms in the grammatical systems Halliday introduces (18) the conceptually unimpeachable but terminologically infelicitous[13] distinction between *tonality*, *tonicity*, and *tone*, often referred to as the "three Ts" by later researchers (e.g., Wells 2006). *Tonality*, somewhat opaquely, is the division of speech into intonational phrases; *tonicity* is the location of the *tonic* or nuclear accent; and *tone* is the pitch choice made for the nucleus (and any tail).

There is a clear statement of the phonological structure of the *tone group* (intonational phrase) (13):

(7) $(\mathrm{P}_{1\ldots n})\mathrm{T}(_{2\ldots n})$ OR $(\mathrm{P}_{1\ldots n})\mathrm{T}(_{2\ldots n})\mathrm{T}(_{2\ldots n})$

The basic structure consists of *pretonic* (P) and *tonic* (T), corresponding to Palmer's head and nucleus. The second alternative accommodates the kind of double or *compound* nucleus already discussed (O'Connor and Arnold's fall + rise compound tune). The subscripts refer to optional feet, and the only obligatory foot is the tonic. As in other accounts, the pitch movement associated with the tonic is distributed over the material after the tonic syllable, and as no contrasts are available post-tonically, no separate element of structure is recognized (14), a view with which I agree. Less persuasive is the lack of a prehead element, though we may presume this to be because it has not been shown to be an exponent of a grammatical system. Also phonological is the recognition of phonotactic restrictions: "There are of course probability restrictions on the combination of tonic and pretonic" (17), some examples of which are given.

There is a fairly conventional inventory of five *primary tones* to which numbers are allocated: 1, *falling* (to low pitch); 2, *rising*, subsuming falling-rising (both rising to high pitch); 3, *rising* (to mid); 4, *(rising-)falling-rising* (to mid); and 5 *(falling-)rising-falling* (to low) (16). Compound nuclei are made up of tones 1 and 3 or 5 and 3. Listing the "terminal tendency" (e.g., to mid) is reminiscent of boundary tones in AM, though evidently with three values in the inventory, but this aspect is not developed. There is then a breakdown, at the next degree of delicacy,[14] into *secondary systems at tonic* and *secondary systems at pretonic*. The latter is basically an inventory of head patterns; while for instance one of two terms relating to tone 5 is a tonic "rising to mid, falling; 'breathy.'" At this point it seems that the phonological analysis is straying into what many would regard as subphonological realizational detail and into paralinguistic uses of phonation quality. It is easy to imagine an incredulous rise-fall-rise with breathy phonation on “Me!?” but that breathiness is equally possible on an incredulous and shocked fall on “No!”—risking an explosion of “secondary systems” of tone.

Unlike the other approaches we have considered, Halliday’s (1967) analysis is clearly theoretically rather than pedagogically oriented. If we take the examples used to illustrate the “tone-expounded [secondary] system” of “commitment” as it applies to clauses in the declarative “mood,” we find notations that require considerable decoding (41):

(8) *Commitment:*
neutral—tone 1 //1 ∧ it’s / possible // (= “I don’t know”)
uncommitted—tone –3 //–3 ∧ it’s / possible // (= “I don’t care”)
committed—tone5 //5 ∧ it’s / possible // (= “so what?”)

// indicates tone group boundaries and / foot boundaries; underlining marks the tonic syllable; and ∧ is the silent *ictus* (salient syllable) of a foot—that is, the pretonic in (8) lacks an initial salient syllable, in a situation that would elsewhere be analyzed as a prehead. Given a pretonic foot, albeit a defective one with no ictus, secondary system 1 means high pitch on the first foot and a fall to low, –3 means a low pretonic and a tonic rising to mid, and 5 means a rising-falling tonic, and puzzlingly no secondary system 5 is specified (17).

From the theoretical point of view, Crystal (1969b, 391–392) sees the fundamental problem in Halliday’s analysis being the attempt to extend the concept of system from grammar, where it applies without tension, to intonation:

> Intonation patterns are on the whole far less discrete formally and semantically, are far less finite, are more difficult to identity, and so on, than grammatical patterns. . . . Intonation “systems,” in short, are NOT “just as much” grammatical . . . as are those of tense, number, etc.: Halliday is too optimistic about the predictability and stability of intonation contrasts. (392)

9.2.3.4 Wells Wells (2006) is a work very much in the pedagogical tradition established by Kingdon and O’Connor and Arnold, and it draws also on work by Halliday. The modern term *intonational phrase* is used, but its structural “anatomy,” to use Wells’s term, is that of O’Connor and Arnold, with an obligatory nucleus, and optional prehead, head, and tail, in all of which positions except the tail significant choices can be made. From Halliday’s work comes the terminology *tonality*, *tonicity*, and *tone* (the “three Ts”).

The exposition concentrates at first exclusively on contrasts arising from the nuclear choice (*tone*), limited in discussion to *fall*, *rise*, and *fall-rise*. Wells’s presentation is organized by grammatical type, with mainly attitudinal attributes of the different intonations being exemplified. For instance, “a wh question can also be said with a non-fall: a rise or, less commonly, a fall-rise. This has the effect of making it *more gentle*, kindly, encouraging, sympathetic or deferential, as opposed to the businesslike fall” (43). Tonicity and tonality are then discussed in chapters 3 and 4, and only then does Wells look “beyond the three Ts.” This is where the elements of the intonational phrase before the nucleus are considered, in terms of prehead, head (the first accent of which is termed *onset*), nucleus, and tail. Wells’s contrasting patterns in the head are *high level*, *high falling*, *low level*, and *low rising* (209). There are complex versions of high level, high falling, and low rising heads, these involving reduplication just as in the stepping, sliding, and climbing heads of O'Connor and Arnold (1973). Mixed complex heads are possible, but they are outside the scope of the book. A marked high prehead can be chosen rather than the fairly low normal prehead (215). A final point of interest to be mentioned here, redolent of Halliday's (1967) secondary systems, is the refinement of nuclear categories in Wells's chapter 5: *high falls* are distinguished from *low falls*; a *rise-fall* is added; fall-rises are separated into *high* and *mid* types, this apparently defined by the end point of the rise and augmented by a *rise-fall-rise*; the *mid-level* nucleus is introduced; and three types of rise are recognized, namely *high*, *low*, and *wide*. It is a little unclear to what extent all these are full members of the nuclear inventory, merely left until later to present for pedagogical expediency, or are instead implementational variants.

The question of what patterns contrast has always taxed the British school, just as it does the AM framework. Without an equivalent to the knockout phonemic question "Is this the same word?" and with intonation multitasking across grammar, attitude, information structure, and other functions, it really is difficult to decide what is worthy of abstraction as a phonological category (cf. Ladd and Morton 1997) and what can be left to gradient phonetic implementation. Authors such as O'Connor and Arnold have presented inventories that were clearly delineated, though not necessarily objectively correct. Wells, despite the generally clear presentation, gives me the impression of having been (understandably) troubled by this issue of what is phonological and of not having fully resolved it for the reader. Although he provides a usefully updated presentation of the British school, in some ways he seems to have missed a chance to move forward from O'Connor and Arnold—perhaps by stripping out pitch range variation more explicitly as a distinct system and operating with a somewhat reduced inventory of nuclei, or in some other way sharpening up the analysis in the light of the dichotomy between categories and gradience.

9.2.3.5 Crystal Leaving Crystal (1969a) to last in this whistle-stop tour through the history of the British school is counter-historical, but the unique status and stature of *Prosodic Systems and Intonation in English* make it appropriate. In the preface Crystal notes that despite existing work on intonation and voice quality, "there is a marked reluctance to develop any theoretical perspective, or to provide criteria for evaluating different partial descriptions" (vii). His aim is to provide that theoretical basis by integrating intonational phenomena within a more general theory of nonsegmental phonology. Untrammeled by the need to deliver a minimal description appropriate for language teaching, he approaches prosody from first principles and develops an impressively comprehensive framework. This is based on a corpus of three hours of recordings of thirty university-educated speakers of southern British English (see also Crystal and Quirk 1964).[15] The analysis, by which Crystal means the end product of a structural description, is achieved by careful listening and auditory interpretation, with recourse to the native speaker judgments of others when needed (8) and occasional reference to acoustic correlates (16).

Crystal (1969a) regards *prosodic* features as constituted by variation in pitch, loudness, duration, and silence (128), while *paralinguistic features* are divided into *voice qualifiers*, such as breathiness and creak, and *voice qualifications*, such as giggle and tremulousness; paralinguistic features lie at the least discrete, least linguistic end of a continuum, and function solely to convey attitude. This means that *prosody* encompasses a wide range of phenomena, including some that are gradient, and Crystal's framework has a degree of complexity that reflects this. He views intonation "not as a single system of contours, levels, etc., but as a complex of features from different prosodic systems. These vary in their relevance, but the most central are *tone*, *pitch-range* and *loudness*, with *rhythmicality* and *loudness* closely related" (195). His is thus what he terms a "parametric" approach, which he acknowledges is at least implicit in previous work by writers such as Sweet and Palmer to the extent that they discuss pitch range as a variable independent of pitch movement. He deals quite extensively with matters such as "tonal subordination" and the grammar and the semantics of intonation, but I will limit my comments to the formal descriptive system.

Crystal agrees with others in the British school on the centrality of the *tone unit* (intonational phrase) and its division into *prehead, head, nucleus*, and *tail*, as well as the importance of nucleus placement. Questionably, he refers to the nucleus as the accentual syllable of the most prominent word (210); as noted in section 9.2.2.3, the notion that the nucleus must be the most prominent element is not always borne out. Where he most clearly innovates in respect of the intonational pitch parameter is in developing a comprehensive mechanism for describing pitch range, applicable to syllables anywhere in the intonational phrase, and thereby removing the dilemma of whether to recognize high and low versions of kinetic tones. His *simple pitch-range system* (144–150) relates the pitch of a syllable to the preceding syllable: by default, stressed syllables are slightly lower (reflecting declination), but they may by choice be somewhat lower again than this default, or very much lower; at the same pitch; or slightly higher, much higher, or very much higher.[16] The complex pitch-range system (150–156) expresses the extent of pitch movement: default, wide, or narrow in the case of the nuclear syllable, with additionally the possibility of *monotone* in polysyllabic sequences. In combination the simple and complex pitch-range systems provide seven times three combinations (or four in the case of polysyllabic sequences), so a paradigm of twenty-one (or twenty-eight) choices (if my understanding and arithmetic are right). With "simple" and "complex" descriptive terms for tempo (e.g., *allegro; rallentando)*, loudness (e.g., *forte, diminuendo*), and systems of rhythmicality (e.g., *rhythmic/arhythmic*, with *spiky/glissando* describing pitch movement between syllables and *staccato/legato* alternations of loudness and duration between syllables), tension, and pause, a framework of apparently unbounded power and subtlety is constructed.

From today's perspective, this seems like abandoning phonological principles in favor of an impressionistic surrogate for quantitative description of phonetic implementation, contrary to Crystal's explicit claim to a "phonological" approach (e.g., 200). However, his use of *phonological* is to contrast his own approach, which he insists puts the identification of prosodic units first, with the alternatives of starting from grammar or semantics. And, given his definition of paralinguistic systems in such a way as to, in effect, exclude those parameters that also engage in signaling at the more linguistic end of the intonational continuum, it is inevitable that his prosodic systems allow phenomena to be described that many would regard as outside the remit of intonational phonology. It may seem that Crystal pushes the methodology of the British school to its limits and possibly beyond. Whether one accepts the validity of the approach, it is undeniable that in its attempt to give a comprehensive account of prosody, facing as it does Janus-like toward core language structure and toward gradient paralinguistic signaling, his analysis and his data highlight many phenomena that cannot be ignored if a comprehensive account of prosody is sought.

9.3 Critical Comparison of the British School and AM

In the middle part of the twentieth century, the main alternative linguistic framework of intonation serving as the “other” against which the British school could be compared was the American levels framework. This described intonation not in terms of pitch movements but in terms of pitch levels, usually four, numbered 1 to 4, but without universal agreement as to which end was highest. Because this tradition is largely superseded, I propose not to deal with it and rather to bring the British framework further into focus by contrasting it with the current paradigm dominant in intonation research, namely the bitonal autosegmental-metrical (AM) framework (see Ladd 2008 for an accessible survey; the label itself for this tradition of work is due to Ladd—cf. Gussenhoven 2004, 123). In this I shall take Pierrehumbert’s (1980) seminal analysis of American English as a fixed reference point, while acknowledging that many variants and modifications have emerged since then, most prominently the slightly simplified tones and break indices (ToBI) version (Silverman et al. 1992; Beckman and Ayers 1994) that aimed to be a practical standard for intonational analysis and has been widely used for English and implemented for other languages. This diversification, however, does not affect the basic differences between the two broad frameworks.

9.3.1 Common Ground between the British School and AM

The British school and those working in the AM framework share an essentially phonological approach. A phonological approach, to remind ourselves, means allowing a degree of abstraction in the search for connections between the acoustic signal and the meanings conveyed by it; not, for instance, expecting direct mapping between a meaning and a physically constant aspect of the signal, such as a particular curvature and excursion of the fundamental frequency contour and, say, the meaning “interrogativity.” Rather, a phonological approach requires the analyst to posit abstract elements—intonational primes such as a rise or fall nucleus—having the potential to be associated with multiple meanings or functions: for instance, a rise may signal a question, but it may also signal sympathetic involvement in a response that requires no answer (e.g., ˈnever ˏmind).

Furthermore, as with segmental phonology, the primes will have variable realizations according to factors in the context in which they occur. We have seen this acknowledged repeatedly by the representatives of the British school discussed in this chapter. The realization is described impressionistically and graphically, but it is unambiguous that the mapping of the intonational prime is sensitive to the context in which it occurs. Ladd (2008, 10), it will be recalled, requires that a fully phonological approach (as opposed to a “proto-phonological” one) be able to state the mapping between the phonological categories and objectively measured physical parameters, but I argued that the relevant domain for mapping for most of the British school era had, perforce, to be auditory impressions, and that in principle the British school adhered fully to a phonological conceptualization of intonation.

AM and the British school both posit an intonational phrase within which coherent melodic patterns can be identified and that is generally regarded as independent from grammatical structure; although the terminology differs, to a first approximation this primary unit is equivalent in the two approaches. AM has grown up in an era when the notion of a “prosodic hierarchy” is prevalent and has correspondingly been more inclined to recognize other phrasing domains, notably the intermediate phrase (ip) following Beckman and Pierrehumbert (1986). However, although most pedagogical works in the British tradition stuck with a single domain, there was an awareness theoretically that other larger or smaller domains, and relations between domains, would be necessary for a full account. Trim (1959, 28) notes that there can be unifying features across two or more melodically well-formed intonational phrases, such as a progressive descent in the pitch of equivalent intonational events, which imply the operation of a higher-level unit, and he proposes the “major tone-group,” which would consist of one or more “minor tone-groups”; for instance:

(9) ˇaid | veri mʌtʃ ˋlaiktuː | bət ˇfɑːðə | wount ˋhiərəvit ||

Crystal (1969a) develops descriptive apparatus for “tonal subordination”: “The primary characteristic of the subordinate tone-unit is that its pitch-contour, while having a complete and independent shape within itself, falls broadly within the total contour presented in the superordinate tone-unit—cf. Trim’s ‘major’ and ‘minor’ tone-groups” (245). While these proposed units may not be coterminous with the intonational phrase (denoted as IP) and ip, there is a shared recognition of the need for a hierarchical structure of intonational units.

Both frameworks agree, explicitly or implicitly, on the importance of the metrical structure of utterances, at least to the level of distinguishing (in one terminology or another) stressed and unstressed syllables. In Steele ([1775] 1969) we see a lively appreciation of the tight relation between melody and stress-based prominence, and Halliday (1967) very explicitly refers to the foot, represented $I(R_{1\ldots n})$—that is, initiated by a “salient” syllable at “ictus” and followed by zero or more weak syllables at “remiss.” The majority of British analyses and notations reflect the notion that the stressed syllable is the element to which the start of a significant pitch movement is anchored. AM also associates pitch accents with metrically strong syllables, so both frameworks recognize the symbiotic relationship between metrical prominence and melodic choices.

9.3.2 Differences between the British School and AM

An immediately striking difference is that the British school uses abstract pitch changes as its primes and AM uses abstract pitch levels or targets. Relatedly, the favored representations are iconic in the British school (sloping lines and/or arrows), and the letters H (high) and L (low) in the case of AM. This associated difference is not inevitable; it would be possible to devise a notation for the AM analysis using, for instance, high and low horizontal lines instead of H and L; however, matters have evolved as they have. It can be argued that the tonetic stress marks developed in the British framework have greater immediacy and are more accessible to learners. Perhaps those brought up with H and L would disagree.

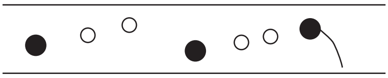
There is superficially persuasive evidence that the target approach has greater empirical validity. It has been shown that if the length of segmental material, for example, a vowel, is varied experimentally, L and H targets show stable pitch height and stable alignment with (or “anchoring to”) segmental or syllabic landmarks, with the result that the slope and duration of the glide joining the targets vary, suggesting that it is the targets and their alignment that are the constants (e.g., Arvaniti, Ladd, and Mennen 1998; Ladd 2008, 172ff). The glide is merely an interpolation between the targets and not the phonological prime.

However, this may not be quite the knockout argument it is usually taken to be. We are quite happy to work with segments such as a voiceless plosive filling slots in structure—for instance, the onset or coda in the syllable. From a phonological point of view, the plosive is atemporal, whether represented as a unit phoneme or as an array of phonological features, but to realize it phonetically, it must receive exponence in terms of independent phonetic parameters aligned in time, so that, for instance, the glottal opening can be aligned appropriately with the oral closure—co-extensive for an unaspirated stop, persisting after release for aspirated stop, and starting early for preaspirated stop. An abstract rise, equivalently, can be seen as entailing a start parameter and a stop parameter, both of which have to be appropriately aligned. In any case, Hs and Ls are not targets but abstractions whose realization is not unmediated but dependent on contextual interpretation (Pierrehumbert 1980), and it is a little inequitable to demand direct realization from abstract falls and rises. A kinetic tone could be regarded, along the lines of Gussenhoven (1984, 197), as an "intonational morpheme" made up of the "purely formal" elements H and L, it being those latter elements that are phonetically aligned.

That said, the case for glides would be strengthened to the extent it can be shown that perception attends to change rather than targets, and even then, the AM theorist could retort that our well-evidenced perceptual reliance on formant transitions in vowels to determine place of articulation has not deposed a target-based segmental phonology. What is clear is that while realization, in principle, is a matter in AM of mapping from phonological categories by rule to physical dimensions,[17] in the British school realization is dealt with by exemplification using impressionistic representations. This, however, is not an inevitable difference but a consequence of the contemporaneous methodology.

Following Pierrehumbert (1980, 29–30), orthodox AM theory regards the intonational phrase as containing (in addition to boundary tones) a sequence of pitch accents of equal status,[18] whereas the British school developed, in broad outline, an internal structure of prehead, head, and nucleus where different inventories could apply. (I omit the tail, which most writers treat as not contrastive independently of the nucleus.) The extent to which different inventories were posited for different places in the structure varies: for instance, Kingdon (1958) allows all the kinetic tones in the head, whereas O'Connor and Arnold (1961, 1973) have a reduced inventory with distinct symbols. This may reflect a tension between completeness of phonetic description and pedagogical expediency. What I think the structural analysis of the intonational phrase achieves, in general, is greater clarity of description of the semantic and pragmatic function of melodies. The naturalness of reduplicated heads, as for instance in (10), also suggests a unity to the head.

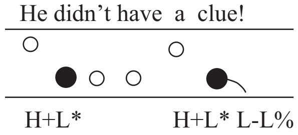
(10) ↗Why did you ↗choose to do `that?



Part of the motivation for the structural analysis of the intonational phrase, however, is the special status of the nucleus. Whole tune descriptions were in effect nuclear analyses, and all analyses regard the nucleus as crucial in characterizing the intonational phrase. This, however, brings us to a separate but related difference, namely the recognition in AM of boundary tones. In this, I think, AM has a pretty watertight case. The British school has tended to flounder a bit trying to explain why nuclei tend to group functionally according to their final pitch direction (e.g., rise, fall-rise, rise-fall-rise), and why in a (rise-)fall-rise you can't always get the implementation done and dusted immediately, but have to plod through a long tail at low pitch before producing an exuberant upward flourish at the last moment. The boundary tone analysis makes sense in both these respects. Boundary tones are not, in principle, incompatible with British-style analyses, as we shall see in section 9.4.

The British school treats pitch accents as left-headed,[19] that is, starting on a metrically strong syllable. Halliday (e.g., 1967) is perhaps the writer who most explicitly links the melodic elements, the tones, to the stress-initial foot (often called "Abercrombian" after, e.g., Abercrombie 1967, 131), but the association is implicit widely across the British framework. In AM models, on the other hand, left-headed pitch accents such as L* + H coexist with right-headed ones such as L + H*, in keeping with the analysis in Bolinger (1958) to whom the term *pitch accent* in this sense can be attributed. On the face of it, the British approach is the more constrained and so constitutes the stronger and more testable theory with fewer ambiguities. For instance, an utterance that could be analyzed in a generic British approach as "↗How's he ˏcoping?" with a rising head and low-rise nucleus, could be L* + H L* L-H% or L* H + L* L-H%, where the first high target is attributed to the first or the second accent. However, there are patterns that challenge the descriptive adequacy of the British models. In (11), for example, it is hard to accommodate the high *a*.

(11)



A high prehead followed by a low head captures the first four syllables, but the second high point is not modelable. As always in intonation it is hard to establish a pattern as contrastive; is the pattern in (11) in fact distinct from the same except for a rising head? But even if the melody shown is a free variant of the rising head, the British analysis less well captures the "reduplicative" nature of the melody than the repeated H + L*. It is a shame that, as a result of the polarization of effort between the two intonational frameworks, such matters of equivalence and descriptive adequacy between different frameworks have not been subject to more investigation.

The freedom to have right-headed pitch accents, and in most variants two boundary tones, allows AM to model a larger number of potentially contrasting pitch accents. One such case is shown in figure 9.4, exemplifying pitch accents on the monosyllable *yes* (ToBI representations and typical names for the nuclei are given below the patterns). The British analysis lacks a nuclear tone corresponding to the second pattern (H + L* L-H%).

Interestingly, however, a little-adopted mechanism in Kingdon (1958) can accommodate the second pattern. He suggests that although a syntagmatic contrast in pitch between a prehead and the head appears to require at least one initial unstressed syllable, the contrast can be effected "by pronouncing the beginning of the stressed syllable on a pitch which will contrast with the tone to be used, and delaying the tone and its accompanying stress for a length of time sufficient to allow the initial unstressed contrasting pitch to be felt" (53). This he terms a "homosyllabic prehead."[20]

Not all lexically stressed syllables are pitch accented (i.e., showing pitch obtrusion). Those that are not, but that receive some degree of prominence, are commonly termed *stressed* (as opposed to accented) in the British tradition and often marked in texts by a

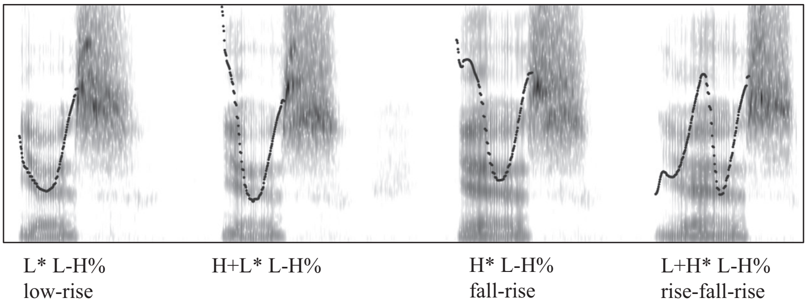
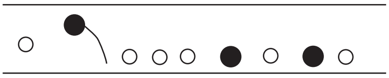


Figure 9.4
Rising nuclei on the word *yes* represented in AM and British notations. The second pattern with a right-headed H + L* pitch accent has no equivalent in mainstream British notation, though it could be accommodated by Kingdon's homosyllabic prehead. The fundamental frequency contour (dots; 50–175 hertz) is shown superimposed on the spectrogram (0–8 kilohertz).

dot or small circle. I suggested (section 9.2.3.2) that a better terminology would be to reserve "stress" for the lexical property of prominence-readiness, while syllables lacking a pitch accent but that are prominent in an utterance might be regarded as being *rhythmically accented*, this term reflecting the timing adjustments that achieve that prominence and implying also cues in intensity and spectral balance. Orthodox AM descriptions have no facility to recognize prominence in the intonational representation[21] other than by pitch accents, which means that in an utterance such as that in (12), with a high fall nucleus near the start marking contrastive prominence on "why," and a long tail, the prominent but low-level syllables "keen" and "sell" would have to be left unmarked or, redundantly and misleadingly (because there is no new pitch movement), be given L* pitch accents—arguably after an L- intermediate phrase tone required to create the sharp nuclear fall. In this respect the British analysis, with rhythmic accents marked by dots, offers the simpler but more comprehensive prosodic interpretation.

(12) But `why are you so · keen to · sell it?



H* L- (L*) (L*) L%

In summary, the hard-won insights of the British school often tally with the concepts and principles of AM, but there are clear divergences of detail. In some cases the British tradition seems better to capture the facts, and in other cases AM holds the winning hand. In the next section, we explore a model that is the offspring of both traditions.

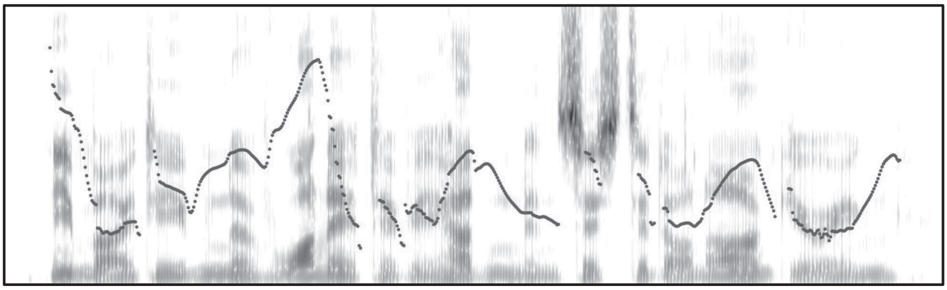
9.4 The British School Meets AM: IViE

IViE (Intonational Variation in English) refers to a collection of audio recordings of young adult speakers of urban varieties of English in the British Isles collected in the years around the millennium (Nolan and Post 2014; Post and Nolan 2012). The name and acronym also refers to a version of the bitonal AM model (Grabe, Nolan, and

Farrar, 1998; Grabe, Post, and Nolan, 2001; Grabe 2001) that is more compatible with the British school than is orthodox AM. IViE draws heavily on the work of Gussenhoven (1984, 1990) who had adapted H/L notation in a way influenced by analyses in the British tradition.

The IViE notation adopts the bitonal H and L mechanism from AM, but by restricting itself to left-headed pitch accents retains a greater degree of interconvertibility with analyses in the British school. On the other hand, the boundary tone insight is adopted in IViE. These features can be seen in figure 9.5, which compares a generic British analysis (panel a) with the equivalent bitonal representation in IViE (panel b) and ToBI (panel c). It can be seen that whereas the ToBI analysis attributes the prenuclear rising and falling patterns to interpolation to a following right-headed pitch accent (H + L* or L + H*), the other two analyses agree on left-headed pitch accents. This requires a high prehead, or equivalently an initial %H boundary tone, to be recognized. Note also that ToBI has two levels of boundary, T% and T-, the latter being the boundary of the lower level "intermediate phrase." IViE only has one, T%; though this could be revised in the light of later empirical evidence from the AM research enterprise and in accord with Trim's (1959) proposal for major and minor tone groups.

Separately, IViE also allows the specification 0%, indicating an intonational phrase boundary where no new pitch movement is evident. Thus the specification of a falling nucleus, with a sharp fall on or shortly after the nuclear syllable, is H* + L 0%,[22] which arguably represents the facts more transparently where a long, low tail follows the nuclear syllables than the ToBI specification of H* L-L% does.[23] There have been objections that 0% violates the bitonal character of AM by introducing a third value, but actually it merely extends the principle to boundaries that not every syllable is marked for pitch. This principle also justifies the practice in IViE, following the British school, of marking syllables that are prominent by means other than pitch obtrusion. These rhythmic accents (see section 9.2.3.2) can be marked by *, with no tonal specification because no new pitch instruction applies, as for instance in the IViE representation in (13) for the example given in (12):



(a) ¯hɪ ↗maɪt iːvn əv `wʌn ɪt | ɪf ɪ ↘hædnt ɪn ↘sɪstɪd ɒn ˇɑː g j ʊ ɪ ŋ ‖

(b) %H L* + _H H* + L 0% H* + _L H* + _L H* + L H%

(c) H + L* H* L- H* L + H* L + H* L- H%

Figure 9.5

A constructed example consisting of two intonational phrases analyzed using (a) a generic British model, (b) IViE, and (c) an orthodox AM model. The fundamental frequency contour (dots; 50–175 hertz) is shown superimposed on the spectrogram (0–8 kilohertz).

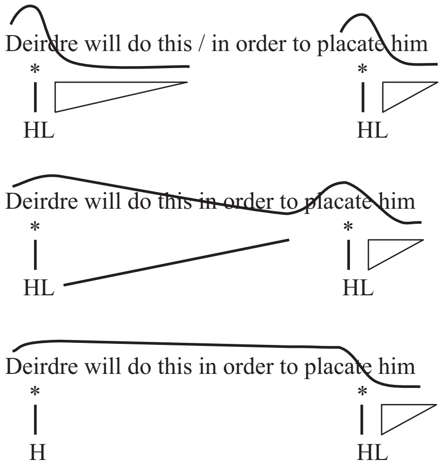
(13) But `why are you so · keen to · sell it?

H*+L * * 0%

In fact, IViE complements its phonological tier where the intonological specifications (such as H* + L) are shown, by a separate *prominence tier*; though this is useful primarily as a first stage in the process of analysis where the crucial prominent syllables are to be identified. Also of largely heuristic value in the IViE scheme is a *phonetic tier*, where candidate pitch accents can be labeled as, for instance, mHm, meaning a high turning point on a prominent syllable preceded and followed by syllables on a mid pitch, or lMh-l, which might characterize a "delayed peak"' falling pitch accent (see Grabe 2001 for more examples).

Gussenhoven (1990, 32–34) proposed that "tonal domains" could undergo restructuring, as shown in his example roughly reproduced here as (14). In this example, the sequence HL | HL |—two nuclei, in effect, in two separate tonal domains—could (for instance under faster speech) become H_L HL | in a single tonal domain (a falling head and falling nucleus, in traditional terms, with the first L reassociating as far right as possible), and even H HL | (a high head plus fall) by "tonal linking," that is, deletion.

(14)


Deirdre will do this / in order to placate him
*
*
HL
HL
Deirdre will do this in order to placate him
*
*
HL
HL
Deirdre will do this in order to placate him
*
*
H
HL

The motivation of this essentially derivational analysis is the perceived "relatedness" of the contours; however, it is not hard to find cases where the phrasing is informationally significant. For instance, the presence of first phrase boundary in "the boys (|) who are bold by nature | are ready to speak up" is crucial for marking the nonrestrictive status of the relative clause, as opposed to its absence for a restrictive reading. Nonetheless, the approach can capture some valid equivalences. The IViE model adopts, in principle, Gussenhoven's notion of restructuring in the sense that the equivalent of a "falling head," H*+_L, is symbolized as a fall with the low target displaced; though in practice the three patterns are usually notated separately as H* + L 0%, H*+_L, and H*.

9.5 Conclusions

I hope to have shown that the British school of intonation analysis in the shape it attained in the mid-twentieth century was the result of incremental development, each step arising from sophisticated introspection and interpretation by scholars from the seventeenth century onward. The heuristic methods were those of its time, and the British school's most consistent motivation was the need to present intonation in a way that would allow it to be taught explicitly to language learners. As a consequence it does not have the quantitative empirical underpinning of the AM paradigm that is dominant today, with its access to rapid acoustic analyses and its potential applications in speech technology. Nevertheless, studies within the British school were probably the first to recognize, implicitly at least, that phonological principles were needed for the analysis of speech melody; this is evident in the attempt to set up inventories of significantly different pitch patterns at different places in structure, patterns that were abstract and allowed of context-sensitive implementation. The success of the British school is attested, I would suggest, by the fact that some of its essential characteristics—its recognition of intonational phonology, and the centrality of metrical prominences as loci for pitch accents, for instance—match those of AM, and in the majority of cases converting a description in one framework into one in the other is eminently feasible.

The wealth of quantitative research that AM stimulated in the last quarter of the twentieth century and subsequently has not been matched within the British school, and indeed several of the research themes, such as alignment, appear to be crucially dependent on the use of H and L pitch targets or turning points. However, I have suggested, albeit tentatively, that analysis of kinetic primes such as rises and falls into component elements is perfectly coherent within the British framework, parallel either to featural analysis of phonological segments or, as Gussenhoven suggests, to the decomposition of morphemes into segments. The real question, then, would be whether the higher unit—the kinetic element—adds any value to the lower level analysis using H and L, either in terms of psychological plausibility or pedagogical advantage. These are questions that I don't think have been answered, and perhaps are too seldom asked, but they should be amenable to empirical study.

From a pedagogical standpoint, one might hazard a guess that the iconic notation of the British school involves less cognitive load for a learner than the autosegmental notation, whereas the generalizations permitted by the AM boundary tone facilitate the connection of form and meaning. If so, there is no barrier to incorporating a boundary tone into an otherwise traditional iconic notation, as in (15) where the utterance is the one shown in figure 9.5:

(15) ¯hɪ ↗maɪt iːvn əv ˋwʌn ɪt | ɪf ɪ ↘hædnt ɪn ↘sɪstɪd ɒn ˋɑː g j ʊ ɪ ŋ ↑|

The final fall-rise is indicated not as a unitary fall-rise but as a fall on the nuclear syllable and a rise associated with the major phrase boundary. I did for a while teach intonation to phonetics students using such a notation, but made the mistake of trying to do so simultaneously with an equivalent IViE notation in terms of H and L—which is probably like learning to use chopsticks and a knife and fork at the same time. My prediction would be that such a modified British notation, taught alone, would actually be easier to master than either the no-boundary-tone traditional version (because it would explicitly show those cases where part of the "nuclear" pattern is delayed), or the AM notation (which requires the H and L elements to be composed into pitch movement), but this can clearly only be tested by pedagogical research.

In summary, the British school of intonation has a history that can be traced back several centuries, and it uncovered many of the central insights into the phonological and phonetic structuring of intonation that are common to other frameworks. Its largely pragmatic, pedagogical orientation, in an age when instrumental analysis of speech was not readily accessible, meant that it did not receive the extensive quantitative validation that AM has triggered; however, its more iconic notation may give it an advantage in its pedagogical function. Its analyses are often compatible with those of AM, and, where they are not, it sometimes has solutions that better do justice to the data, such as the recognition of syllables accented by rhythm alone (often termed "stressed" in opposition to "[pitch-]accented"). On the other hand, further development and application of the British approach to intonation analysis would do well to incorporate insights adopted by AM, such as the boundary tone.

Notes

1. Spelling has been normalized here from the part-phonetic orthography used by Butler. Pages refer to the 1910 reprint.

2. Biblical quotations are used as examples, for instance: *asking*—"How oft shall my brother sin against me, and I forgive him?" (Matthew 18, 21); *urging/reprehension*—"Are ye so without understanding also?" (Mark 7, 18); *avowing the contrary*—"Can the blind lead the blind?" (Luke 6, 39).

3. The debate originally took place in correspondence. Lord Monboddo was won over and graciously accepted Steele's position—see Abercrombie (1965, 37).

4. I shall use the currently widely accepted *intonational phrase* neutrally for the main unit within which intonational patterns are discussed, even though it is anachronistic with respect to most of the work discussed; it allows reference to the concept without having to choose between different terms used by particular authors of the British school such as "breath group," "tone group," "tone unit," and "intonation group."

5. In retrospect, the choice of the term *head* is not ideal, given its current use in linguistics as, roughly speaking, the defining and obligatory element in a construction. Palmer could scarcely have foreseen this usage. In intonation, it is the nucleus that defines the nature of the intonational phrase, with the head potentially unfilled and if filled playing a subsidiary role in the flavor of the intonational phrase.

6. I have replaced the segmental phonetic transcription of the original with orthography and approximated the graphical representation.

7. This devil is, however, conjured by those authors who, for functional reasons, define the nucleus as being "the most prominent" syllable rather than the last pitch-accented syllable; see Ladd (1982) and Nolan (1984). My students tend not to agree that a downstepped falling nucleus after a high head in an utterance such as "ˈreally ˎnice" is the more prominent element. Probably writers have conflated "most prominent" with "most significant for the semantics of the phrase."

8. It will, however, be noted that in the "broken" scandent head reproduced in example (5), the reduplications begin reliably on lexically stressed syllables, so there is an implicit recognition of the bond between pitch pattern and metrical prominence.

9. In that it apparently reduplicates the nuclear position.

10. This is reminiscent of Kingdon's more open combination of kinetic tones. It might have been an infelicitous attempt to accommodate what is simply a more emphatically spoken version of the sliding head or a reluctance to recognize the need for a second, lower level of intonational phrasing allowing multiple nuclear falls in succession.

11. Attitudinally different, in their account; most saliently, in my view, the fall-rise prepares for a "but"—e.g. "but you're wrong to think I live in Sheffield too," whereas the fall + rise introduces a new topic ("mother") while picking up an old topic (Sheffield), without any implied reservation.

12. Not defined, however.

13. From the point of view of remembering and distinguishing them.

14. I hope I have this usage correct within the systemic framework.

15. They are speakers of RP or near-RP, which Crystal describes as "geographically unmarked." Actually it isn't. Although it may serve as a prestige norm throughout England and to a lesser extent elsewhere in the United Kingdom, it is part of the sociophonetic continuum of the south-east of England, and RP speakers will by default be assumed to be from there.

16. In this last case the "high booster" syllable will be higher than a previous high pitch prominent syllable.

17. Though note interesting exceptions such as the work of Jun on Korean (e.g., 2005, 208–210) which distinguishes a contrastive *phonological* tone tier and a *phonetic* tone tier where the melodies are neither contrastive nor fully predictable.

18. Though as Ladd (2008, 133–134) notes, this position is widely relaxed in practice.

19. This is an example of why the British school's use of *head* for the optional, subordinate, prenuclear element in the structure of the intonational phrase isn't ideal.

20. Classics purists might prefer "tautosyllabic."

21. The "metrical" part of the analysis would represent prominence relations, but this is rarely shown explicitly.

22. The + and the 0 have commonly been omitted; here, for example, H*L %, but at some risk of misreading—in this case as H* L%, which would imply a gradual, not sharp fall if the postnuclear material were longer.

23. H* + L, in orthodox AM following Pierrehumbert (1980), was not available to mark surface falls but was used to trigger downstep.

References

Abercrombie, D. 1965. *Studies in Phonetics and Linguistics*. London: Oxford University Press.

Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Armstrong, L. E., and I. C. Ward. 1926. *A Handbook of English Intonation*. Leipzig: Teubner. Second edition (1931) and subsequent editions, Cambridge: Heffer.

Arvaniti, A., D. R. Ladd, and I. Mennen. 1998. "Stability of Tonal Alignment: The Case of Greek Prenuclear Accents." *Journal of Phonetics* 26:3–25.

Beckman, M., and G. M. Ayers. 1994. *Guidelines for ToBI Labelling*. http://www.speech.cs.cmu.edu/tobi/.

Beckman, M., and J. Pierrehumbert. 1986. "Intonational Structure in English and Japanese." In *Phonology Yearbook 3*, edited by Colin Ewen and John Anderson, 255–309. Cambridge: Cambridge University Press.

Bolinger, D. 1958. "A Theory of Pitch Accent in English." *Word* 14 (2–3): 109–149.

Brown, G., K. L. Currie, and J. Kenworthy. 1980. *Questions of Intonation*. London: Croom Helm.

Burnett, J. 1784. *Of the Origin and Progress of Language*, Vol. IV. Edinburgh; London: Bell; Cadell.

Butler, C. (1633) 1910. *The English Grammar, or, the Institution of Letters, Syllables, and Woords in the English Tung*. Edited by A. Tichler. Oxford: Turner.

Chomsky, N., and M. Halle. 1968. *The Sound Pattern of English*. New York: Harper.

Coleman, H. O. 1914. "Intonation and Emphasis." *Miscellanea Phonetica* 1:6–26.

Cruttenden, A. 1990. "The Origins of Nucleus." *Journal of the International Phonetic Association* 20:1–9.

Cruttenden, A. 1997. *Intonation*. 2nd ed. Cambridge: Cambridge University Press.

Crystal, D. 1969a. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Crystal, D. 1969b. "Review of Halliday, M. A. K. 'Intonation and Grammar in British English.'" *Language* 45 (2): 378–393.

Crystal, D., and R. Quirk. 1964. *Systems of Prosodic and Paralinguistic Features in English*. The Hague, the Netherlands: Mouton.

Danielson, B., ed. 1955. *Works on English Orthography and Pronunciation: 1551, 1569, 1570. Part I, Biographical and Bibliographical Introductions. Texts and Index Verborum*. Stockholm: Almqvist and Wiksell.

Eichler, A., ed. 1910. *Charles Butler's English Grammar*. Halle, Germany: Niemeyer.

Grabe, E. 2001. "The IViE Labelling Guide." Version 3. http://www.phon.ox.ac.uk/~esther/ivyweb/guide.html.

Grabe, E., F. Nolan, and K. Farrar. 1998. "IViE: A Comparative Transcription System for Intonational Variation in English." In *Proceedings of the Fifth International Conference on Spoken Language Processing*, Vol. 4, edited by R. H. Mannell and J. Robert-Ribes, 1259–1262. Sydney, Australia: Australian Speech Science and Technology Association.

Grabe, E., B. Post, and W. F. Nolan. 2001. "Modelling Intonational Variation in English: The IViE System." In *Prosody 2000*, edited by S. Puppel and G. Demenko, 51–57. Poznan, Poland: Adam Mickiewitz University.

Gussenhoven, C. 1984. *On the Grammar and Semantics of Sentence Accents*. Dordrecht, the Netherlands: Foris.

Gussenhoven, C. 1990. "Tonal Association Domains and the Prosodic Hierarchy in English." In *Studies in the Pronunciation of English*, edited by S. Ramsaran, 27–37. London: Routledge.

Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Halliday, M. A. K. 1967. *Intonation and Grammar in British English*. The Hague, the Netherlands: Mouton.

Hart, J. 1551. *Of the Opening of the Unreasonable Writing of Our Inglish Toung*. In Danielsson 1955.

Hart, J. 1569. *An Orthographie, Conteyning the Due Order and Reason, How to Write or Paint Thimage of Mannes Voice, Most Like to the Life of Nature*. London: Serres.

Jones, D. 1909. *Intonation Curves*. Leipzig, Germany: Teubner.

Jones, D. 1918. *An Outline of English Phonetics*. 1st ed. Leipzig, Germany: Teubner. Subsequent editions, Cambridge: Heffer, and Cambridge: Cambridge University Press.

Jun, S-A. 2005. "Korean Intonational Phonology and Prosodic Transcription." In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S-A. Jun, 201–229. Oxford: Oxford University Press.

Kingdon, R. 1958. *The Groundwork of English Intonation*. London: Longman.

Ladd, D. R. 1982. "Review of Brown, G., Currie, K. L. and Kenworthy, J. 'Questions of Intonation.'" *Language* 58:204–208.

Ladd, D. R. 2008. *Intonational Phonology*. 2nd ed. Cambridge: Cambridge University Press.

Ladd, D. R., and R. Morton. 1997. "The Perception of Intonational Emphasis: Continuous or Categorical?" *Journal of Phonetics* 25:313–342.

Nolan, F. 1984. *Auditory and Instrumental Analysis of Intonation*. Cambridge Papers in Phonetics and Experimental Linguistics, 3. Cambridge: Department of Linguistics, University of Cambridge.

Nolan, F., and B. Post. 2014. "The IViE Corpus." In *The Oxford Handbook of Corpus Phonology*, edited by J. Durand, U. Gut, and G. Kristoffersen, 475–485. Oxford: Oxford University Press.

O'Connor, J. D., and G. F. Arnold. 1961. *Intonation of Colloquial English*. London: Longman.

O'Connor, J. D., and G. F. Arnold. 1973. *Intonation of Colloquial English*. 2nd ed. London: Longman.

Palmer, H. E. 1922. *English Intonation, with Systematic Exercises*. Cambridge: Heffer.

Palmer, H. E. 1933. *A New Classification of English Tones*. Tokyo: Institute for Research in English Teaching.

Pierrehumbert, J. B. 1980. "The Phonology and Phonetics of English Intonation." PhD diss., MIT.

Post, B., and F. Nolan. 2012. "Data Collection for Prosodic Analysis of Continuous Speech and Dialectal Variation." In *The Oxford Handbook of Laboratory Phonology*, edited by A. C. Cohn, C. Fougeron, and M. Huffman, 538–547. Oxford: Oxford University Press.

Silverman, K., M. E. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, T. Kohler, and J. Wesener Hirschberg. 1992. "ToBI: A Standard for Labeling English Prosody." In *Proceedings of the Second International Conference on Spoken Language Processing*, 867–870. Banff, Canada: International Speech Communication Association.

Steele, J. (1775) 1969. *An Essay towards Establishing the Melody and Measure of Speech, to Be Expressed and Perpetuated By Certain Symbols*. Facsimile ed. Menston, UK: Scolar Press.

Sweet, H. (1877) 2013. *A Handbook of Phonetics: Including a Popular Exposition of the Principles of Spelling Reform*. Oxford: Clarendon. Published digitally by Cambridge University Press (2013), http://dx.doi.org/10.1017/CBO9781139854665.

Sweet, H. 1892. *A Primer of Phonetics*. Oxford: Clarendon.

Trim, J. L. M. 1959. "Major and Minor Tone-Groups in English." *Le Maître Phonétique* 112:26–29.

Wells, J. C. 2006. *English Intonation: An Introduction*. Cambridge: Cambridge University Press.

10 The PaIntE Model of Intonation

Antje Schweitzer, Bernd Möbius, Gregor Möhler, and Grzegorz Dogil

10.1 Introduction

The PaIntE model (Möhler and Conkie 1998; Möhler 2001) was originally developed for F0 modeling in text-to-speech (TTS) synthesis. Its purpose was to generate F0 contours that are as close to natural, human-produced F0 contours as possible. We will show in this chapter that it can also be used for more general research on intonation. The PaIntE model assumes that only the F0 contour in the vicinity of so-called intonation events contributes to the intonational meaning of an utterance, whereas the stretches in between these events arise from interpolation and do not affect the overall meaning. This view is manifested in the model's name: parameterized intonation events, or PaIntE for short.

PaIntE can be classified as a sequential model of intonation, in that it composes the F0 contour from a sequence of local contours, each associated with some kind of meaningful tonal event, and that these events or local contours do not interact or affect each other. PaInte shares this view with the well-known, phonologically motivated, tone sequence model (TSM; Pierrehumbert 1980), for instance. The sequential approach is also popular in speech synthesis, with the Tilt model (Taylor 1998) as probably the most widespread example.

While the PaIntE model has long been successfully used to generate perceptually appropriate F0 contours in speech synthesis, we will show in this chapter that PaIntE is also well suited for intonation research beyond speech synthesis.

To this end, we will introduce the PaIntE model in some detail in section 10.2, focusing on several aspects. First, we will show that PaIntE shares assumptions with autosegmental models of intonation (section 10.2.1). Second, in contrast to many other models, which aim to identify typical shapes that correspond to intonation categories, PaIntE assumes that several dimensions contribute to the meaning of tonal events and quantifies these dimensions by continuous parameters. The overall shape of the event is then determined by these parameters, as illustrated in some detail in section 10.2.2. Next, we provide more information on how the PaIntE parameters can be derived from a database (section 10.2.3) or how they can be predicted for speech synthesis (section 10.2.4).

In section 10.3, we will illustrate that the PaIntE intonation events can be related to categories posited by autosegmental approaches to intonation. Also, as we will show in section 10.4, PaIntE can be used for answering typical questions in intonation research. To this end, we will discuss some recent studies that have used the PaIntE model to investigate intonation, some from an autosegmental perspective, and some from an exemplar-theoretic perspective. Thus, sections 10.3 and 10.4 show that PaIntE is compatible with an autosegmental approach to intonation, but it can also serve an exemplar-theoretic approach. Its flexibility lies in that, on the one hand, it can take autosegmental categories into account, and that the PaIntE parameters can be linked to these categories. On the other hand, however, it does not necessarily assume that such categories exist. We discuss the advantages of this property of PaIntE in section 10.5 and offer some conclusions in section 10.6.

10.2 The PaIntE Model and Its Parameters

To motivate the requirements and objectives of an F0 model in the context of speech synthesis, we briefly sketch the role of such a model in the TTS synthesis process. We then consider the commonalities between intonation models for synthesis and more general models of intonation from a theoretical point of view, before turning to the specific implementation of PaIntE in terms of its parameters and how these parameters can be extracted for analysis, or predicted for synthesis.

10.2.1 Intonation Models and Speech Synthesis

Traditional concatenative TTS systems generate speech starting out from a given, text-only specification of the utterance to be synthesized. This specification is passed through a pipeline of mostly independent modules, each of which incrementally adds linguistic, phonological, and phonetic information. Toward the end of this process, an F0 model adds concrete F0 values to the specification, and at the end of the pipeline, synthesized speech is generated by concatenating speech segments from a recorded database and, if necessary, manipulating these segments to match the F0 values that the F0 model has predicted.

What F0 models for synthesis and more general intonation models have in common, at the very least, is that they are interested in relating aspects of meaning to tonal contours. For instance, a TTS system might want to relate sentence-internal major syntactic phrase boundaries to rising intonation contours and sentence-final syntactic phrase boundaries to falling contours. Or it may relate exponents of information structure to contours conveying the intended meaning.

Many concatenative TTS systems treat the problem of predicting F0 contours following what may be called a phonological approach. According to Ladd (1996, 11), a phonological model of intonation has to minimally consist of two ingredients: first, a finite set of intonation categories, and second, a mapping from these categories to continuous acoustic parameters. In this vein, many TTS systems take into account linguistic properties inferred from the text to first predict the occurrence of a finite set of intonation categories such as pitch accents or boundary tones and then generate concrete F0 values in a second step. In fact, in the context of TTS, the term *F0 model* often refers just to this second step, that is, the generation of an F0 contour given a specification that already contains the desired location of some kind of intonation categories in the utterance.

In the case of a phonological approach to the TTS problem as just outlined, an additional commonality between F0 models for synthesis and more general intonation models is that both types of models have to address two issues: (i) identifying the relevant categories and (ii) specifying how these categories are implemented phonetically in terms of F0 (and, realistically, in terms of other prosodic cues such as duration).

A TTS system that follows such a phonological approach has the additional objective of specifying a mapping from linguistic properties to intonation categories. This is not necessarily an objective of a more general model of intonation. However, to establish a distinction between two intonation categories, even a more general model would have to show that exchanging the two categories in some utterance context changes the meaning of the utterance.[1] Therefore, a more general intonation model cannot be entirely silent regarding the relation between meaning and intonation categories. The exact nature of this relation is still an open issue. For instance, there is consensus that in West Germanic languages such as English and German, information structure affects pitch accent placement (Terken and Hirschberg 1994; Féry and Kügler 2008) and even the type of pitch accent (Pierrehumbert and Hirschberg 1990; Chen, den Os, and de Ruiter 2007). However, to capture this impact, researchers have to refer to fine distinctions in information status that naive speakers are probably not aware of. Pierrehumbert and Hirschberg (1990), for example, elaborate differences among five meanings, termed *new*, *addition of new value*, *accessible*, *modification of given*, and *given* by Baumann, Röhr, and Grice (2015), that are claimed to give rise to different pitch accents. However, Baumann and Grice (2006) show that for German, a more fine-grained notion of accessibility is needed because accent types differ depending on the way in which the accessible information can be inferred from the text. Similarly, Baumann, Röhr, and Grice (2015) differentiate ten classes of information status that differ in accentedness and type. A crucial difference, then, is that a TTS system can rely only on more coarse-grained meaning that can be estimated from raw text, namely, text without annotation or markup, because this is what serves as input to a TTS system. Even worse, the TTS system has to expect that the estimated meaning may be incorrect at times.

Given these difficulties, most TTS systems treat the mapping from linguistic properties to intonation categories as a separate task, which is addressed as part of the linguistic analysis of the text to be synthesized (Sproat 1998; Taylor 2009). Then the task of the F0 model is “only” to map from a specification of the utterance, which already includes intonation categories, to concrete intonation contours. In other terms, the task of the F0 model in synthesis is to provide the phonetic implementation of phonological categories that have been determined in a preceding linguistic analysis step. This separate treatment is fostered by the dissemination of speech corpora with manually annotated tones and break indices (ToBI) labels (e.g., Ostendorf, Price, and Shattuck-Hufnagel 1996; Rapp 1998; Calhoun et al. 2010; Eckart, Riester, and Schweitzer 2012), which conveniently serve as training and testing data for this second step in F0 modeling.

Arguably, the separate treatment may also reflect a split in research approaches between studies that relate meaning to intonation categories (e.g., Beckman 1996; Büring 1997; Féry 1993; Pierrehumbert and Hirschberg 1990; among many others) and studies that investigate phonetic detail in the implementation of F0 contours (e.g., Pierrehumbert 1981; Kohler 1990; Ladd, Mennen, and Schepman 2000; van Santen and Möbius 2000; among many others).

The PaIntE model follows the practice of separating the prediction of categories from the actual F0 modeling. It does not necessarily state what the exact nature of the categories is. All that is said is that they are “intonation events,” and the core task of the PaIntE model then is to generate concrete F0 contours that implement these intonation events. However, PaIntE acknowledges that ToBI categories are an obvious and convenient choice in that respect and provides means to take ToBI categories as the relevant intonation events for which local F0 contours have to be generated, by way of a configuration parameter. Before we go into detail on how the PaIntE parameters of a given contour can be extracted, and how the parameters can be predicted in speech synthesis, we first discuss how they determine the concrete F0 shapes.

10.2.2 The PaIntE Parameters

All intonation models dedicated to F0 modeling for speech synthesis parameterize the F0 contour in some way. In the case of PaIntE, the shape of the F0 contour around local intonation events is captured by six linguistically motivated parameters. Together they determine the F0 contour in a window of up to three syllables centered on the event. In the original implementation, the events were taken to be pitch accents and boundary tones posited by a German GToBI variant (Mayer 1995), that is, the PaIntE parameters served to specify the exact shape of the F0 contour on and around pitch accents and boundary tones. The global contour then arises by interpolation between these events.

Mathematically PaIntE employs a function of time, with $f(x)$ giving the F0 values at time x. It is defined as follows:

$$f(x) = d - \frac{c_1}{1 + e^{-a_1(b-x)+\gamma}} - \frac{c_2}{1 + e^{-a_2(x-b)+\gamma}} \tag{10.1}$$

This function yields a peak shape (figure 10.1), where the first term, the d constant, gives the upper bound. We will see that d can be interpreted as *peak height* parameter. From this d constant, two sigmoids are subtracted—the second and third terms in the equation. The first of these two sigmoids alone would result in a falling shape, starting at c_1 in negative infinity ($\lim_{x\to-\infty} = c_1$) and ending at 0 in infinity ($\lim_{x\to\infty} = 0$). The most pronounced part of this fall ends approximately at the value for parameter b. Because this sigmoid is subtracted from the d constant, this effectively yields a rise toward d, that is, toward the peak height parameter. The amplitude of this rise is c_1, and the pronounced part of the rise ends approximately at the value for parameter b. In the same way, subtracting the second, originally rising, sigmoid adds a fall component to this rising shape. The pronounced part of the fall starts close to parameter b, that is, approximately at the point where the first sigmoid levels off. Thus, we get a pronounced peak with the peak location affected by the b parameter; in other words, b can be interpreted as the *peak alignment* parameter. As parameter d is the upper bound for rise and fall, and thus the upper bound for the peak, this parameter corresponds to peak height. The amplitudes of the rising and falling parts are determined by parameters c_1 (rise amplitude) and c_2 (fall amplitude), and their steepness by parameters a_1 (steepness of rise) and a_2 (steepness of fall).

PaIntE provides several methods to normalize the time axis. In the standard variant, which is called *sylnorm* normalization, the time axis inside the approximation window is normalized such that syllable boundaries occur at integer values, with the syllable related to the intonation event beginning at 0 and ending at 1. In the sylnorm case, b determines the temporal alignment of the peak in terms of relative position within the syllables in the approximation window. A hypothetical example peak contour for a syllable associated with a pitch accent, using sylnorm normalization, is given in figure 10.1. The PaIntE function as specified in equation 10.1 is indicated by the solid line. Syllable boundaries are indicated by vertical lines, the syllables themselves are indicated by σ symbols, and the pitch-accented syllable is marked as σ*. The location of the b parameter (peak alignment) is marked by the bold vertical line; parameters c_1 (rise amplitude) and c_2 (fall amplitude) are indicated by the arrows, and the d parameter (peak height) by the tick at the y-axis. Parameters a_1 (steepness of rise) and a_2 (steepness of fall) cannot be read off the graphical representation in the same way as the other parameters, but they are hinted at in figure 10.1.

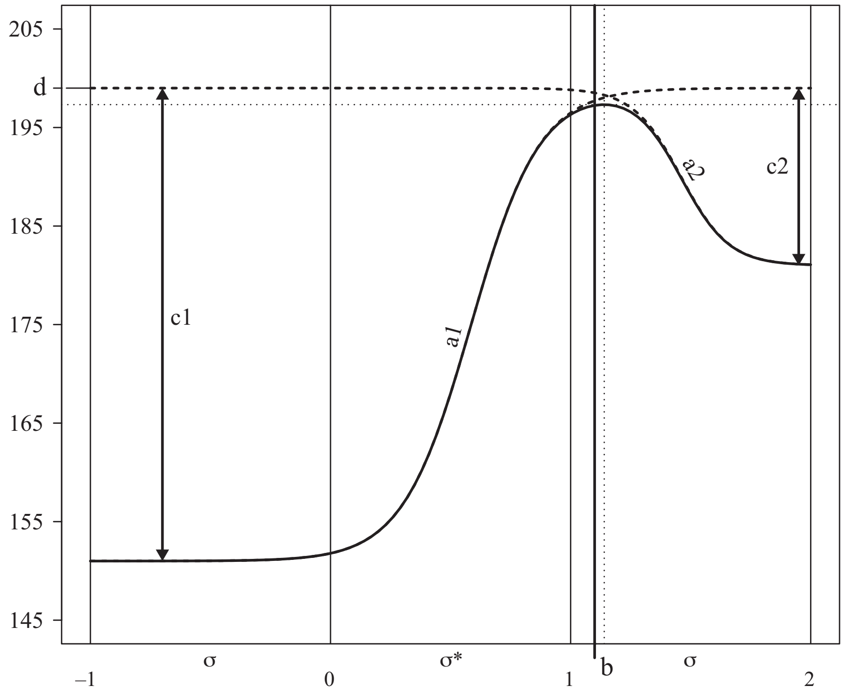


Figure 10.1
Example PaIntE contour in a window of three syllables around a pitch-accented syllable (σ*). See the text for details.

The shape with the pronounced peak is not prototypical for all syllables that are associated with some intonation event. Often, we observe just a falling or just a rising contour, without a clear peak. To accommodate such cases when parameterizing existing F0 contours, the PaIntE model first tries to detect a peak in the three-syllable window. It uses the PaIntE function as specified in equation (10.1) only in case there is a peak. If no peak is detected, the function is used with only the first sigmoid for rising contours or only the second sigmoid in case of falling contours, yielding the following two functions:

$$f_{rise}(x) = d - \frac{c_1}{1 + e^{-a_1(b-x)+\gamma}} \tag{10.2}$$

$$f_{fall}(x) = d - \frac{c_2}{1 + e^{-a_2(x-b)+\gamma}} \tag{10.3}$$

We have indicated these alternative functions by dashed lines in figure 10.1. They are partly hidden by the PaIntE function (solid line), but it can be seen that the *b* parameter, indicated by the bold vertical line, does indeed occur at the point where the dashed line for the rising sigmoid starts to level off, and where the fall in the dashed line for the falling sigmoid is about to become more pronounced. It is thus a reasonable estimate of the end of the rise, given that, mathematically, the rise does not end at all because it never reaches *d* before infinity. Similarly, the fall effectively starts in (negative) infinity, not somewhere in the window depicted here. The same argument can be made for parameter *d*; it is only an estimate for the height of F0 at the end of the rise, or the beginning of the fall, respectively, and is never really reached. However, it can be seen in figure 10.1 that for the two example sigmoids, the difference between *d* and the function values at the edges of the three-syllable window is not detectable by eye; the dashed line for the falling sigmoid seems to reach the *y*-axis exactly where the tick for *d* is located.

It should also be noted here that in the case of the "full" PaIntE function with the peak as given in equation (10.1), the *b* parameter is again only an approximation of the temporal alignment of the peak in the syllable structure. It can be seen in figure 10.1 that the peak's exact temporal location, as indicated by the dotted line, is actually a small distance to the right of the *b* parameter itself, which is indicated by the bold vertical line. The exact displacement depends on the values of the a_1, a_2, c_1, and c_2 parameters. In this specific example, for instance, the true peak is 0.04 units to the right of the *b* value, that is, if the last syllable in this example were two hundred milliseconds long, *b* as an approximation of the temporal alignment of the peak would be off by eight milliseconds. If a better estimate of the peak alignment is desired, one can resort to sampling the curve specified by the PaIntE function and finding the point where the maximum sample occurs.[2]

To get an impression of the accuracy of *b* for approximating the location of the peak covering a representative number of contexts, we have estimated both *b* and the true temporal alignment of the peak for approximately seventeen thousand syllables in a database of two hours of speech that had been approximated using the full PaIntE function in equation (10.1). The mean absolute error in syllable units was approximately 0.052 and the median 0.039. In absolute time, this corresponded to a mean absolute error of approximately ten milliseconds and a median of approximately eight milliseconds. Because the time resolution in deriving F0 from the speech signal is usually in this order, we consider the approximation to be exact enough. However, if necessary, it is easy to estimate the true location of the peak given the six PaIntE parameters as suggested here.

The same point can be made for the *d* parameter (peak height). In the example in figure 10.1, the *d* parameter is 199 hertz, while the true peak height is at approximately 197.3 hertz. For the seventeen thousand syllables from our database, the mean absolute error was approximately 1.380 hertz, and the median absolute error was approximately 1.167 hertz. This is on the order of the just noticeable difference of approximately one hertz in human perception of complex tones at pitch levels below five hundred hertz (Kollmeier, Brand, and Meyer 2008). This leads us to conclude that the approximation is accurate enough for almost all purposes. If a higher accuracy is needed, we again recommend estimating the true peak height numerically by sampling.

So far we have discussed only the application of PaIntE using the sylnorm normalization; however, PaIntE also provides an alternative called *anchor* normalization. In this case, each syllable is split into three parts representing the (unvoiced) onset of the syllable; its sonorant nucleus, which is defined as containing the nucleus and possibly preceding voiced consonants in the onset; and, finally, the coda. This normalization is motivated by findings that indicate that timing in F0 movements is relative to syllable structure (House 1996, 1997; van Santen and Möbius 2000). Using anchor normalization, each syllable in the approximation window is again normalized to length one with the same values for syllable boundaries as in the sylnorm case. Syllable-internally, the unvoiced onset is adjusted linearly to a length of 30 percent of the syllable duration. The sonorant nucleus then spans another 50 percent: it ranges from 30 to 80 percent of the syllable duration, and the coda finally goes from 80 to 100 percent of the syllable duration.

It should be noted that the approximation window may be shorter than three syllables depending on the context. This is the case if silent intervals intervene between the syllables. Also, as stated, PaIntE can be configured to take information regarding prosodic categories associated with the syllables into account. In this case, the window does not extend to neighboring syllables that also carry a pitch accent, or across syllable boundaries that are associated with a phrase boundary. Reducing the approximation window in these cases is motivated by the fact that speakers before boundaries compress tonal contours that in another context would extend into the following syllables (Mayer 1995; Grabe 1998; Jilka, Möhler, and Dogil 1999). Similarly, in cases where it is known that a syllable is associated with a pitch accent that exhibits a late peak (e.g., in an L*H accent as assumed by Mayer 1995), the approximation window does not contain the preceding syllable. However, it is possible to configure PaIntE to enforce the three-syllable window in all contexts, just as it is possible to parameterize every single syllable in cases where no prosodic annotation is available that indicates where intonation events are expected.

10.2.3 Extracting the PaIntE Parameters

Before the PaIntE parameters of a given F0 contour can be approximated, the raw F0 contour is smoothed to eliminate micro-prosodic effects and outliers. To this end, PaIntE uses the smooth_f0 algorithm based on the Edinburgh speech tools (Taylor et al. 1999). Smooth_f0 is a median smoother that interpolates across unvoiced regions, but not across silences.

In preparing the approximation, PaIntE first looks for an F0 peak in the smoothed contour in the middle of the approximation window, as well as local minima to the right and left of the maximum. The locations of the maximum and the minima, as well as the number of frames available, are used to determine which of three approximation methods is appropriate in that particular context:

- *Mean F0 approximation*. No PaIntE approximation takes place if there are fewer than two voiced frames for the current window or if the two minima are less than five frames apart. In these cases, PaIntE reverts to a simple approximation called *meanf0* by just determining the mean F0 value in that window as the *d* parameter (peak height); the five other PaIntE parameters are set to 0.
- *Single sigmoid approximation*. If either the left or the right minimum coincides with the maximum, that is, a rise or fall has been detected, but a clear peak has not, the PaIntE approximation is modified to leave out one of the two sigmoids, as described in section 10.2.2. Depending on which sigmoid is left out, this is called the *rise_sigmoid* or the *fall_sigmoid* method. In this case, the *a* parameter (steepness of rise or fall) of the missing sigmoid is set to –1, and its *c* parameter (rise or fall amplitude) is set to 0. The remaining parameters are determined by the single sigmoid approximation.
- *PaIntE approximation*. In the standard case, in which a peak has been detected, that is, neither minimum coincides with the maximum, the approximation is carried out using the PaIntE function as defined in equation (10.1). This is called the *pfun* method.

The approximation itself determines the PaIntE parameters using the appropriate functions, choosing the parameters so that the root mean squared error (RMSE) between actual F0 values and the corresponding values in the PaIntE function is minimized. Finding the optimal combination of parameters is an optimization problem, and PaIntE uses a conjugate gradient method to arrive at a local optimum.

10.2.4 Predicting the PaIntE Parameters

There are several ways in which the PaIntE parameters can be used to generate F0 contours in speech synthesis. As discussed, speech synthesis systems often approach predicting F0 contours as a two-step problem, in which first a set of intonation categories, for instance, the ToBI categories, is predicted from the text. As a result, one would have a specification of the utterance to be synthesized that already includes concrete ToBI categories. The task of F0 modeling can then be viewed as mapping from the ToBI category to the PaIntE parameters, taking linguistic and phonological context into account.

Typically, this mapping would be learned using machine learning techniques on a database that is annotated with all the context properties that will be available at synthesis time, and with the category labels. Most machine learning schemes predict only one parameter at a time; thus, one approach would be to train six models, each of which predicts one PaIntE parameter given the context. However, even if each predicted parameter may be plausible, this does not ensure that the combination of the six predicted parameters is plausible too. One way to avoid this problem is to first determine a finite number of “typical” combinations using clustering techniques. For instance, Möhler and Conkie (1998) used vector quantization, experimenting with between four and thirty-two clusters, whereas Möhler (2001) used up to sixty-four clusters. Then the actual F0 modeling consists of mapping from the given context to one of the clusters instead of to the continuous parameters, thereby turning the regression problem into a classification problem.

The idea of determining a number of typical phonetic implementations by clustering may lead to the question of whether the ToBI categories could be completely replaced by clusters found in this way. However, Möhler (2001) found that the results are better if the GToBI categories proposed by Mayer (1995) are taken into account, both in terms of RMSE between synthesized contour and original, and in terms of correlation between the two. This finding is just one indication that there is a correlation between ToBI-like categories and the PaIntE parameters. In the following section, we show that there is indeed a systematic relationship between these categories and the PaIntE parameters, as the parameters reflect properties that are related to the defining characteristics of the ToBI categories.

10.3 Relating PaIntE to Prosodic Categories

The PaIntE parameters can be related to established categories in a straightforward way. To demonstrate this, we will show here that the PaIntE parameters reflect the expected shape of contours associated with ToBI-style categories—in our case, the tonal categories assumed by the GToBI variant proposed by Mayer (1995). We refer to this variant as German ToBI (Stuttgart variant), or GToBI(S). To this end, we examine PaIntE parameters of pitch-accented syllables extracted from a large database of German read speech.

The database that we use for this purpose was recorded for unit selection speech synthesis (Barbisch et al. 2007) in the course of the SmartWeb project (Wahlster 2004). It was read by a professional male speaker of Standard German and contains typical, isolated utterances of five different genres, usually consisting of one, or at most two, short sentences, corresponding to several prosodic phrases. All utterances were annotated on the segment, syllable, and word level, and prosodically labeled according to GToBI(S). Prosodic labeling for each utterance was carried out by one of three human labelers, supervised and instructed by the first author, in the process of building a database for unit selection speech synthesis. The database amounts to two hours of speech, containing seventy-two thousand segments, twenty-eight thousand syllables, and fourteen thousand words.

GToBI(S) assumes five basic types of pitch accents: L*H, H*L, L*HL, HH*L, and H*M, sometimes described as rise, fall, rise-fall, early peak, and stylized contour, respectively. Just like other ToBI approaches, it assumes that pitch accents are characterized by either a high (H) or a low (L) target associated with the accented syllable and indicates this association using the starred tone notation, H* or L*. It also assumes that the contour before and after the starred tones is determined by trailing and leading tones. In contrast to many other ToBI variants, the notation for these tones is without a + sign to separate trailing and leading tones from the starred tone, but this is a purely notational difference.

Another, less trivial, difference is that GToBI(S) allows tritonal accents: the L*HL accent has two trailing tones, H and L, and the HH*L accent has both a leading H and a trailing L tone. Also, it assumes that the L*H accent and the H*L accent have slightly less prominent allotonic variants, that is, alternative realizations that do not change the underlying meaning of the accent. Mayer (1995) suggests that they can be realized by just the starred tone on the accented syllable, and that the trailing tone can be split off and realized later (partial linking), or can even be omitted completely (complete linking). This results in two monotonal accents, H* and L*, that are interpreted as variants of the H*L and L*H accents, respectively, which differ only in perceived prominence, but not in meaning. Another important difference to the widespread GToBI labeling scheme proposed by Grice and colleagues (Grice and Baumann 2002; Grice, Baumann, and Benzmüller 2005) is that GToBI(S) does not distinguish between an L + H* and an L* + H pitch accent. GToBI(S) provides only the latter category, namely, L*H in Mayer's notation. Cases where other GToBI variants assume L + H* are accounted for in other ways, for instance, by assuming a monotonal H*, where the low pitch level just before the accented syllable is caused by other factors, for instance, by partial linking of a preceding accent.

Apart from this, the expected shape of the pitch contour for each accent is manifested in its notation, as in all ToBI dialects. The contour is expected to reach the target for the starred tone ideally in the middle of the accented syllable, the targets for leading tones on the preaccented syllable, and the targets for the trailing tones on the postaccented syllable or syllables.

Figure 10.2 shows parameterization results for nine pitch-accented syllables selected for illustration from the above-mentioned database. We find that the properties expected given the GToBI(S) categories are reflected in the concrete contours, although there are some differences in the detailed implementation. For instance, in the first accent, identified as an L*HL accent, the rise already starts at the beginning of the accented syllable, reaches the peak at the boundary to the following syllable, and falls within the postaccented syllable. It is thus realized in a more compressed way than expected given its description by Mayer (1995). In the middle panel of the first row, the peak in the H*L accent is at the boundary between the accented and the postaccented syllable, that is, later than the middle of the accented syllable. In the next three

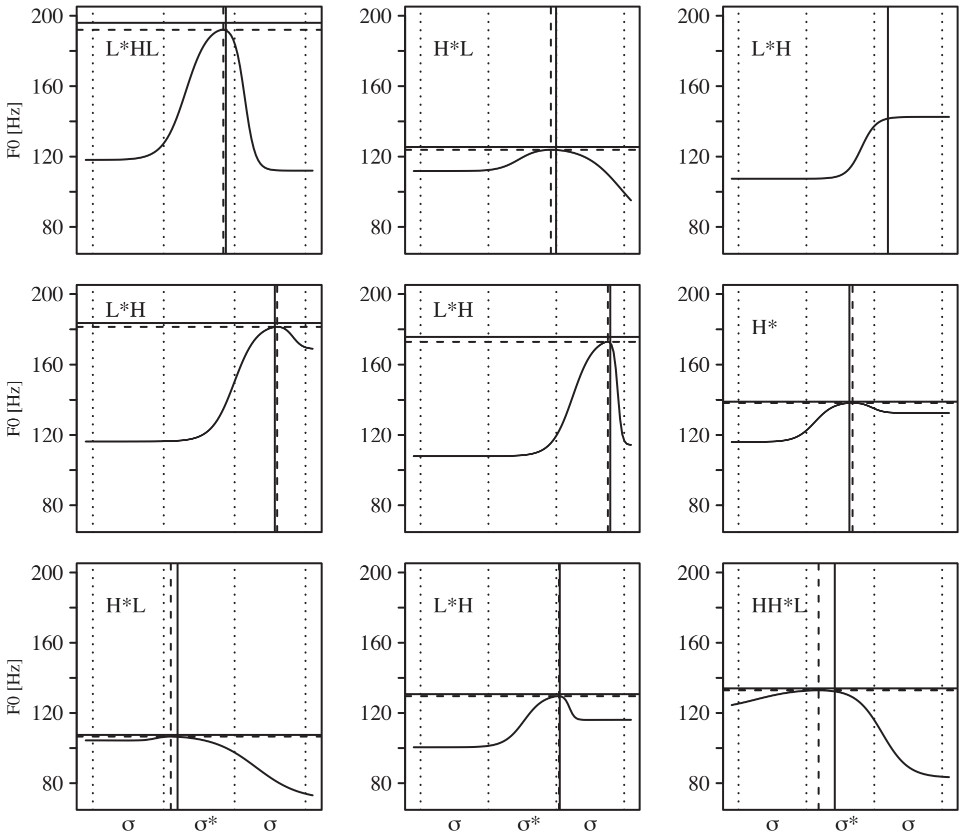


Figure 10.2
PaIntE contours for some GToBI(S) pitch accents. Dotted lines indicate syllable boundaries, σ symbols represent syllables, and the accented syllable is indicated by asterisks. Solid vertical lines indicate the peak alignment parameter b, solid horizontal lines indicate the peak height parameter d, and dashed lines indicate the "true" peak location as estimated using the sampling method outlined in the text. The type of pitch accent according to manual prosodic annotation is indicated within each panel.

accents, all L*H accents, the rise starts on the accented syllable and continues into the next syllable, reaching the peak early in this syllable (in the right panel in the first row), well within this syllable (in the left-most panel in the middle row), or late in this syllable (middle panel in the middle row). Note that in this last example, the contour exhibits a pronounced peak with an abrupt fall, which could be due to a reset to a low pitch level for the next target and thus does not necessarily reflect a property of the L*H accent in question. It should also be noted that the first of these three L*H accents does not exhibit a peak: it was parameterized using the function with only the rising sigmoid.

In the following panel (the right-most panel in the middle row), we can see a prototypical example of a monotonal H* accent, with a less pronounced, broad peak, and relatively low amplitudes, which corresponds well to its characterization as being less prominent than the bitonal H*L variant. Similarly, the three accents in the bottom row reflect the expected properties well: the H*L has a high target in the accented syllable, the L*H rises to a peak at the boundary of the next syllable, and in the HH*L, the contour is already high throughout the preceding syllable and falls to a low level in the postaccented syllable.

To return once more to the question of how accurate the *b* and *d* parameters are, we have indicated the "true" values for peak height and peak alignment by dashed lines in all cases where the approximation was carried out using the full PaIntE function. It can be seen that sometimes there is a small gap between the estimated peak height (solid horizontal line) and the "true" peak height (dashed horizontal line), as evidenced in the top left panel and in the first two panels in the middle row. However, relative to the corresponding rise and fall amplitudes, the difference is small. Similarly, the "true" temporal alignment, indicated by the dashed vertical line, is occasionally at some distance to the *b* parameter as specified in the PaIntE function, indicated by the solid vertical line. This is most obvious in the top middle panel and in the bottom left panel, as well as in the bottom right panel. However, as can be seen in these examples, these larger differences always occur in cases with broader peaks, where one could argue that it is hard to tell where exactly the peak should lie anyway.

These examples demonstrate that it is possible to find the expected properties of pitch accents reflected in the PaIntE parameterization results. But does this also hold on a larger scale, across many examples? To investigate this question, we examine density plots of parameters *b* (peak alignment), c_1 (rise amplitude), c_2 (fall amplitude), and *d* (peak height), obtained from the above-mentioned database. Because GToBI does not make any predictions about the steepness of the contours, we will not discuss parameters a_1 (steepness of rise) and a_2 (steepness of fall) here. Because of limitations in space, we will address only the most frequent of the basic GToBI(S) accents: the L*H, H*L, and L*HL accents.[3]

Figure 10.3 shows density plots of the *b* parameter (*peak alignment*). They are based on parameterization results of the approximately 3,200 syllables with L*H accents, approximately 1,800 syllables with H*L accents, and approximately 270 syllables with L*HL accents in our database. Density plots show how likely a certain range of values is in the underlying data: peaks appear at values that are more likely to occur for the underlying sample, whereas valleys appear at values that are less likely to occur.

We again indicate syllable boundaries by vertical lines. Thus, the broad peak in the dashed line for H*L accents in the left panel, between the two vertical lines, indicates that H*L accents are most likely to have their peak in the middle of the accented syllable. For L*HL accents, which are indicated by the dot-dashed line, surprisingly, the peak is also on the accented syllable. From the description of L*HL accents given by Féry (1993, 94) and Mayer (1995), one would expect this peak to be on the postaccented syllable in many cases. Nevertheless, compared to H*L and H*, the peak for L*HL accents is shifted further toward the syllable boundary. It is also slightly narrower than the two peaks of the H*L and H* distributions, indicating less variation of *b* for L*HL accents.

Finally, the density for L*H accents (solid line) is bimodal: L*H accents are almost equally likely to have their peak either right before the syllable boundary, just as L*HL accents did, or in the later part of the postaccented syllable. One might interpret this as evidence for two distinct categories L + H* (with the peak on the accented syllable) and L* + H (with the peak on the postaccented syllable) as in the more widespread GToBI variant (Grice and Baumann 2002; Grice, Baumann, and Benzmüller 2005); however, the right panel in figure 10.3 shows that this bimodal distribution comes about because L*H is realized differently on word-final syllables than on nonfinal syllables.[4] Here, the dashed line represents L*H accents that occurred on word-final syllables, and the dot-dashed line represents L*H accents that occurred on word-internal syllables. Obviously, these two contexts cause the bimodal distribution: L*H accents on word-final syllables

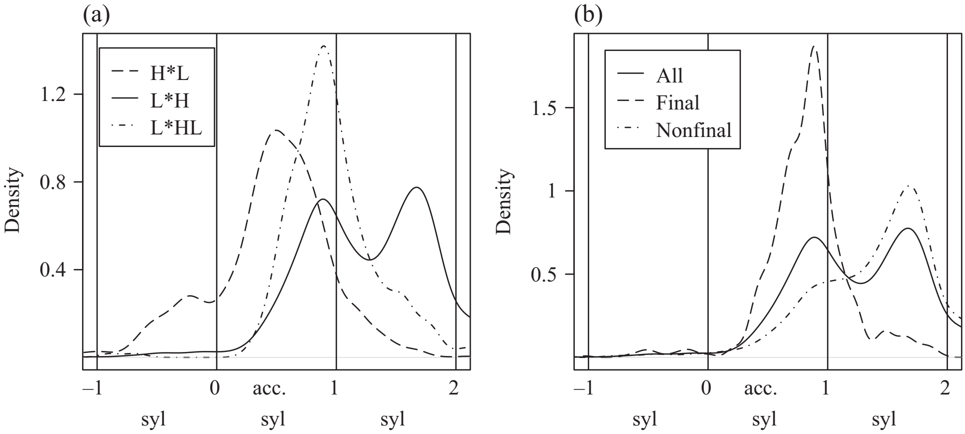


Figure 10.3
Density plots of the b parameter (peak alignment) for some GToBI(S) accents. (Left) H*L accents (dashed line) have their peak earlier in the accented syllable than L*HL accents (dot-dashed line). L*H accents (solid line) have their peak on either the accented or the postaccented syllable. (Right) The bimodal distribution for L*H accents (solid line, repeated from left panel, different scaling) obviously arises because L*H accents in word-final syllables (dashed line) have their peak on the accented syllable, while L*H accents in word-internal syllables (dot-dashed line) have their peak on the following syllable. acc, accented; syl, syllable.

almost always have their peak in the accented syllable, while word-internal L*H accents tend to have their peak on the postaccented syllable. In other words, the tonal movement on L*H accents usually does not cross word boundaries; instead, it is timed to occur earlier before word boundaries.

The example of the alignment parameter *b* shows that the PaIntE parameters not only capture well-known properties of the GToBI(S) accents, but that they can also serve to investigate context-dependent aspects of phonetic implementation, as in the case of word-internal versus word-final L*H accents.

Figure 10.4 gives density plots for parameter *d* (peak height). Values for *d* in hertz are indicated on the *x*-axis: values to the right indicate higher peaks. The figure thus shows that syllables associated with H*L accents (dashed line) are very likely to exhibit low values for *d*: the peak in the line indicates that values just below 120 hertz for peak height are most probable. Compared to L*H (solid line) and L*HL (dot-dashed line), the peaks of H*L accents are lower. This is due to the prevalence of nuclear, that is, phrase-final, H*L accents over prenuclear H*L accents: ninety-three percent of the H*L accents in the database are nuclear accents. For nuclear accents, lower peaks must be expected because of F0 declination, that is, the global trend of F0 to decline over the course of the utterance (e.g., Cooper and Sorensen 1981; Gussenhoven and Rietveld 1988; Pierrehumbert 1979). Indeed, the density plot for non-nuclear H*L accents (not depicted here) is shifted to the right and is broader, similar to the distribution for L*H accents. Peaks of L*HL accents (dot-dashed line) are high even though they are usually nuclear accents in our data (84 percent). The distribution for L*H accents (solid line) is similar to that of L*HL accents (dot-dashed line), although there is again more variation for L*H accents.

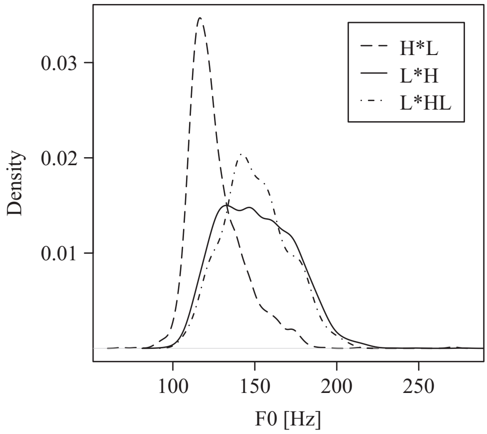


Figure 10.4
Density plots of the *d* parameter (peak height) for the most frequent accents and for unaccented syllables. H*L accents (dashed line) exhibit the lowest values for *d*; L*HL accents (dot-dashed line) and L*H accents (solid line) are characterized by higher and more variable values for *d*.

The last two PaIntE parameters to be discussed here, c_1 and c_2, determine the amplitude of the rise toward the peak (c_1) and the amplitude of the fall after the peak (c_2) in hertz. Figure 10.5 shows the distributions of c_1 (left panel) and c_2 (right panel) for different accent types. Looking at H*L accents first, which are indicated by the dashed line, there is little surprise. It is obvious that they tend to have low values of c_1 (rise amplitude) but higher values of c_2 (fall amplitude): their c_1 distribution shows a pronounced peak for c_1 values of around zero to ten hertz, and although the distribution extends to the right with values of c_1 up to sixty to eighty hertz, the higher values are much less likely. Their c_2 distribution, however, shows a clear dominance of moderately high c_2 values with values of around zero being rather improbable. There is a broad peak between twenty and twenty hertz, indicating that these are typical values of c_2 for H*L accents. In short, H*L accents have small rise amplitudes but higher fall amplitudes, as expected for falling accents. L*H accents (solid line) show just the opposite behavior: their c_1 values are typically between twenty and sixty Hz, while their c_2 values tend to be close to zero, as one would expect for rising accents. The distributions for L*HL accents are given by the dot-dashed lines. They exhibit higher values for both c_1 and c_2, reflecting their characterization as rise-fall accents.

Thus, we have shown for the PaIntE parameters *b* (peak alignment), *d* (peak height), c_1 (rise amplitude), and c_2 (fall amplitude) that their distributions differ depending on which GToBI(S) pitch accent they are associated with, and that the PaIntE model captures the tonal characteristics of the pitch accents well. As a direct consequence of its versatility and accuracy, as well as its linguistic underpinnings, the PaIntE model has recently been employed in a number of intonation studies, which we detail next.

10.4 PaIntE in Intonation Research

In this section, we present several case studies to illustrate PaIntE's potential for intonation research. Moreover, we demonstrate that intonation modeling by means of

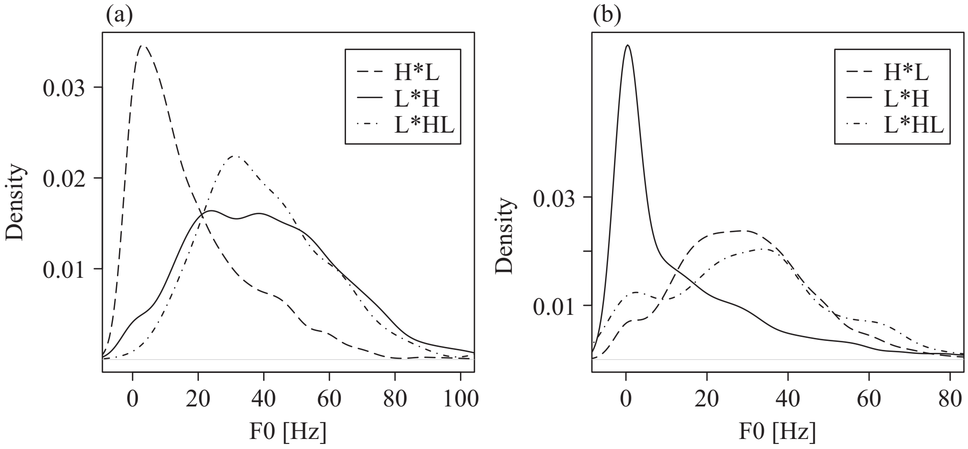


Figure 10.5
Density plots of the c_1 parameter (rise amplitude, left panel) and the c_2 parameter (fall amplitude, right panel) for the most frequent accents. H*L accents (dashed line) usually exhibit low c_1 and high c_2 values; L*H accents (solid line) exhibit low c_2 and high c_1 values. L*HL accents (dot-dashed line) are characterized by high values of both c_1 and c_2.

PaIntE can subserve both the autosegmental-metrical approach (section 10.4.1) and an exemplar-theoretic approach to intonation research (section 10.4.2).

10.4.1 Autosegmental Case Studies

We have shown that general rules concerning the phonetic implementation of intonation categories can be detected in the distributions of the PaIntE parameters as in the case of peak alignment in L*H accents discussed in section 10.3. This methodology can also be employed to test more general hypotheses regarding the implementation of F0 contours. For instance, Dogil and Schweitzer (2011) investigated the alignment of F0 peaks in several German and English databases in this way. Their hypothesis was that there is a quantal effect in peak alignment that causes speakers to place F0 peaks either before or after syllable onsets, but not within onsets. This hypothesis can be motivated by House's (1996) model of tonal perception, which claims that tonal contours within onsets are perceived differently from contours in the nucleus or coda, or by the observation that syllable onsets are considered "weightless" (Goedemans 1998).

To investigate whether peaks in F0 systematically avoid syllable onsets, Dogil and Schweitzer (2011) modified the PaIntE anchor normalization method described in section 10.2.2. Originally, using this normalization, in each syllable, the unvoiced onset is adjusted linearly to take up the first 30 percent of the syllable duration, the voiced onset together with the nucleus to range from 30 to 80 percent, and the coda to span the remaining 20 percent of the syllable duration. In the modified version, voiced and unvoiced onset consonants were treated the same, that is, the onset, regardless of whether it was voiced or unvoiced, was always mapped to the first 30 percent of the syllable duration.

Using this modified normalization, Dogil and Schweitzer (2011) extracted PaIntE parameters from the above-mentioned unit selection database, as well as from a very similar database of a female speaker. Both databases had been manually prosodically labeled according to GToBI(S). The density plots for all pitch accents in these databases exhibited valleys within syllable onsets in their distributions for the *b* parameter (peak alignment), that is, both speakers avoided placing the peak within onsets. The same procedure was applied to a part of the Switchboard corpus of telephone conversations between nonprofessional speakers (Godfrey, Holliman, and McDaniel 1992), which was annotated for accent location (Calhoun et al. 2010), and to German audiobook recordings from the Librivox project,[5] for which no prosodic annotations were available. The valleys were also present in the onsets when looking at the distributions of just the accented syllables from Switchboard, and when looking at the distributions of all syllables in the audiobook corpus, irrespective of whether they were accented.

Investigating PaIntE parameter distributions is not the only way to carry out intonation research using the PaIntE model. It is also possible to investigate the parameters in a more direct way, for instance, by fitting linear mixed models to find which factors affect the PaIntE parameters. For instance, Kelly and Schweitzer (2015) used PaIntE to investigate lexical accents in Trøndersk, a dialect spoken around Trondheim in central Norway. Norwegian distinguishes two lexical accents, named accent 1 and accent 2, respectively. Previous research had found that in Trøndersk, the two accents have a similar shape with a high target followed by a low target, but that they differ in the alignment of tones with the segmental string, with a later timing for accent 2 (Kristoffersen 2006). Also, accent 2 had been shown to have a higher F0 minimum than accent 1 (Kelly and Smiljanić 2014).

Using PaIntE, Kelly and Schweitzer could confirm the previous findings on the later timing of the peak in accent 2: they found that a linear mixed model predicting syllable-normalized parameter *b* (peak alignment) with accent type as a fixed effect was significantly better than the corresponding model without accent type, that is, peak alignment depends on accent type. The study also provided new findings, that is, that accent 2 has a higher F0 maximum than accent 1, and that the amplitude of the fall is smaller in accent 2, again by showing that parameters *d* (peak height) and c_2 (fall amplitude) depend on accent type.

Kelly and Schweitzer's study is, to our knowledge, the first to use PaIntE to investigate lexical accents. The results demonstrate that the PaIntE parameters can be used to assess aspects of phonetic implementation of lexical accents yielding observations comparable to "classical" implementation studies that measure F0 maxima, minima, or turning points. The advantage of the PaIntE model is that these measurements can be derived automatically. This facilitates the investigation of intonation using data on a much larger scale.

10.4.2 PaIntE in Exemplar-Theoretic Approaches

We have demonstrated in section 10.3 that the PaIntE parameters are compatible with an autosegmental view of intonation in that they can serve to specify, or investigate, detailed context-dependent phonetic implementations of ToBI-like categories. This is in fact what PaIntE was designed for originally. Here we will show that PaIntE can also subserve an exemplar-theoretic account of intonation. We briefly introduce the ideas behind examplar theory, and then we give examples of research that has used PaIntE to this end.

In recent years, exemplar theory has gained increasing attention, especially in the segmental domain. The key idea in exemplar theory as applied to speech (e.g., Lacerda 1995; Goldinger 1996, 1997, 1998; Johnson 1997; Pierrehumbert 2001, 2003) is that speakers have access to memory traces (*exemplars*) of previously perceived instances of speech in which almost full phonetic detail is retained. Categorizing new instances in speech perception is based on the stored exemplars and their categories (Lacerda 1995; Johnson 1997; Pierrehumbert 2001, 2003); in speech production, production targets are derived from stored exemplars (Pierrehumbert 2001, 2003).

Under an exemplar-theoretic account, phonetic categories are instantiated by accumulations of similar exemplars in memory. It is sometimes claimed that exemplar models negate abstraction in speech production and perception, but this is not the case. The difference from abstractionist models is that exemplar models assume that abstraction arises as a consequence of generalizing over a large set of exemplars (Pierrehumbert 2003). The aggregation of many exemplars with fine phonetic detail implicitly yields a more abstract linguistic concept with all the properties that the exemplars have in common, leaving all the details in which they vary underspecified. Often it is even explicitly assumed that the exemplars contain category labels (Johnson 1997; Pierrehumbert 2001; Walsh et al. 2010; Wade et al. 2010).

Few studies have looked at prosody in an exemplar-theoretic framework. However, it was shown already in one of the first exemplar-theoretic studies (Goldinger 1997) that exemplars seem to retain prosodic detail, in addition to segmental phonetic properties. In shadowing experiments, subjects tended to adapt their pitch from their baseline pitch toward the pitch of the stimulus token, and to match the durations of their productions to the stimulus token. The effect was stronger for low-frequency words. This indicates that pitch and duration are stored with the word exemplars and that these properties are retained in production.

Building on this, Schweitzer (2011) suggested that even more fine-grained prosodic properties, such as peak height, peak alignment, or rise and fall amplitudes, as quantified by the PaIntE parameters, might be stored in memory. In this vein, Calhoun and Schweitzer (2012) proposed that words and short phrases in American English are stored with their intonation contours and that discourse meanings of highly frequent word-contour pairings can spread by analogy to less frequent pairings. To substantiate this claim, they used PaIntE to parameterize the contours and calculated duration *z*-scores for the segments. Representing the contours by attributes derived from these parameters, they identified fifteen "typical" contours using clustering techniques. They found that certain words and contours form collocations, that is, they appeared together more often than would be expected based on their individual frequencies, supporting the hypothesis that words are stored together with their contours. In a perception experiment, they then confirmed that the discourse meanings of the most frequent pairings spread to other word-contour pairings, which constitutes evidence that the contours were indeed lexicalized.

Further evidence for exemplar storage of prosodic properties comes from a series of three experiments (Schweitzer et al. 2015) demonstrating that phonetic implementation of pitch accents, again in terms of PaIntE parameters, is subject to frequency of occurrence of the linguistic context. We address only the first of these three experiments in more detail here, as it investigates accent implementation in terms of PaIntE parameters directly and thus can serve to illustrate how the PaIntE parameters could be interpreted as dimensions in storing intonation contours. The other two experiments also utilize the PaIntE parameters; however, they are used to assess frequency effects on the similarity or dissimilarity of accent shapes.

The experiment described by Schweitzer et al. (2015) uses a database that was manually annotated for GToBI(S) pitch accents (Mayer 1995). Using generalized linear mixed models, the authors show that accent range in L*H and H*L accents, as quantified by

PaIntE parameters c_1 and c_2, respectively, is significantly affected by the frequency with which the accent and the specific word co-occur. Traditional autosegmental models of intonation, which assume that intonation is postlexical, cannot easily account for such frequency effects, while exemplar models offer a parsimonious account. It is assumed that in production, a number of exemplars that match the required target best are activated and that speakers average over these exemplars, or randomly sample from them, to arrive at a concrete production target. Thus, if a pitch-accented word is to be produced, and if sufficient pitch-accented instances of this word are stored in the speaker's memory, the derived target will match those exemplars and is thus expected to exhibit an F0 amplitude that is appropriate for pitch-accented words. If, on the other hand, however, only a few pitch-accented exemplars of this word are stored, other, nonaccented exemplars will contribute in deriving the production target, leading to a reduced F0 amplitude.

In summary, we would like to argue here that in an exemplar-theoretic account of intonation, the detailed intonational properties that are assumed to be stored with each exemplar can be captured by the PaIntE parameters. We do not necessarily advocate an exemplar-theoretic approach to intonation, but we would like to note that given the problems with labeler consistency and human labeling time, which will be discussed in the following section, an exemplar account does have a certain appeal. However, it should be noted that at least in the second study discussed here (Schweitzer et al. 2015), it is still assumed that exemplars might be labeled with concrete intonation category labels, that is, it does not make such categories obsolete.

10.5 Discussion

In this section, we discuss several theoretical and practical problems of intonation modeling arising from the assumption that the intonation structure of utterances can be described in terms of a linear sequence of intonation events that represent intonational categories. We then move on to discuss the characteristics of PaIntE that allow for a mapping of F0 contours to established intonational categories, but also for analyzing and generating F0 contours in a scenario in which one prefers to remain agnostic with respect to the validity of such categories.

Autosegmental models of intonation aim at establishing a set of intonation categories that, analogously to phonemes in the segmental domain, serve to distinguish meaning. This idea has driven most intonation research in the past fifty years or so (e.g., Goldsmith 1976; Bruce 1977; Gussenhoven 1984; Ladd 1996), with the TSM (Pierrehumbert 1980) and its extension to the ToBI labeling system for American English (Jun, chapter 4, this volume) as one of its most prominent and probably most widely accepted approaches. However, the categories proposed by these models are far from being as established as their segmental counterparts. Even models that do agree on the autosegmental approach differ in the specific inventory of categories that they suggest. In the case of American English, Dilley and Brown (2005), for instance, propose a set of categories for American English that differs from that of the TSM, or ToBI. Similarly, for German, a number of models have been proposed in the autosegmental tradition, all of which assume different category inventories (Kohler 1991; Féry 1993; Mayer 1995; Grice, Baumann, and Benzmüller 2005; Peters 2014). The ToDI transcription system for Dutch intonation (Gussenhoven 2005) focuses on the transcription of tones and limits the number of boundary categories to two phrase boundaries, namely utterance and intonation, without accounting for different strengths of boundaries (unlike ToBI).

Setting aside the problem of agreeing on one authoritative set of categories, a further problem is that even if the categories are taken as given, it is not easy to unambiguously identify these categories in speech data, as evidenced by the moderate consistency with which human labelers can identify them. For instance, regarding labeler consistency for ToBI pitch accents in spontaneous data from the Switchboard corpus (Godfrey, Holliman, and McDaniel 1992), Yoon et al. (2004) report a kappa coefficient of $\kappa \approx 0.51$ for interannotator agreement on the type and presence of pitch accents.[6] However, these moderate consistencies are achieved only when collapsing ToBI pitch accents into two broad categories H* and L*, plus a class X* for uncertain cases; the consistency for the original ToBI inventory must be expected to be even lower. Another study (Syrdal and McGory 2000) on read speech using the original ToBI inventory reports more promising values of $\kappa \approx 0.67$ for ToBI pitch accents in a male corpus and of $\kappa \approx 0.69$ in a female corpus, indicating substantial, but far from perfect,[7] agreement. These corpora, however, consist of read speech by professional newscasters, which has been claimed to be easier to annotate than more spontaneous speech (Mayo, Aylett, and Ladd 1997, 234), and they were annotated by trained and experienced transcribers only.

While newer studies report the kappa coefficient to quantify between-labeler consistency, the first systematic evaluation for ToBI assesses consistency in terms of percentage of matching transcriber-pair-words (Pitrelli, Beckman, and Hirschberg 1994). To calculate the percentage of matching transcriber-pair-words, they carried out pairwise comparisons for each word and each transcriber, accumulating the number of cases where any two transcribers agreed on a particular label (or nonlabel) for a word, and finally dividing this number by the total number of pairs where transcribers either agreed or disagreed. Using this measure, they reported 68.3 percent consistency for pitch accents, while Syrdal and McGory (2000) report 71 percent for their female corpus and 72 percent for their male corpus. Similar values of 71 percent were reported for GToBI pitch accents in German speech data (Grice et al. 1996) and slightly lower values for GlaToBI pitch accents in spontaneous Glaswegian English with 62 percent for nonexpert labelers and 69 percent for expert labelers (Mayo, Aylett, and Ladd 1997). Yoon et al. (2004) report a higher percentage of 86.57 percent, but this again refers to the consistency in labeling their reduced set of pitch accents. Insufficient transcriber reliability was the motivation for the development of ToBI Lite (Syrdal et al. 2000), reducing the set of pitch accent categories in American English to two (rising versus falling), which also served as the basis for the automatic recognition of these categories with high accuracy (the actual intertranscriber consistency was not reported).

Both kappa coefficients and transcriber/word pair accuracies suggest that intonation categories are more elusive than the categories in the segmental domain, where kappa values above 90 percent are not unusual (e.g., Gut and Bayerl 2004). In addition, to make use of these categories in intonation research, sufficiently large databases need to be available that are annotated accordingly. However, manual annotation of these categories is extremely time-consuming. Syrdal et al. (2001), for instance, found that experienced labelers take one hundred to two hundred times real time for annotating ToBI labels.

The fact that human labeling of intonation categories is both time-consuming and prone to labeler inconsistencies makes automatic labeling of these categories all the more attractive. One of the most interesting aspects of the PaIntE model is that it can actually be used to tackle this issue. Schweitzer and Möbius (2009) used the PaIntE parameters to predict pitch accents and boundary tones. They obtained accuracies of approximately 78 percent in the annotation of pitch accent types, and accuracies of approximately 86 percent when addressing the annotation as a two-class problem, that is, when predicting presence versus absence of pitch accent rather than type of pitch accent. These results are slightly but probably not significantly better than those reported for read data by other recent studies (e.g., Hasegawa-Johnson et al. 2005; Sridhar, Bangalore, and Narayanan 2008; Rosenberg 2009).[8]

Unfortunately it is not valid to directly compare the accuracies reported herein to human labeler consistencies in terms of percentage of correct transcriber/word pairs, and studies on automatic labeling do not usually provide kappa values. However, to give an impression of the expected values, we used the Weka tool kit (Witten and Frank 2005) to calculate kappa values for the results reported in Schweitzer and Möbius (2009), obtaining $\kappa \approx 0.62$ for presence or absence and type of pitch accent. This indicates that there was a better consistency between the automatically predicted labels and the human gold standard labels than between human labelers in the study by Yoon et al. (2004), who reported $\kappa \approx 0.51$, but a lower consistency than that reported by Syrdal and McGory (2000) for experts' labels, which was $\kappa \approx 0.67$ and $\kappa \approx 0.69$ in a male and a female corpus, respectively.[9]

To conclude this section, we have argued here that the categories assumed by phonological models of intonation are more elusive than the categories in the segmental domain. Thus, two advantages of the PaIntE model are, first, that it does not depend on the assumption of such categories. Instead, the PaIntE parameters allow for quantifying established parameters such as peak height, peak alignment, or rise and fall amplitudes, on a continuous scale. This can also be exploited under an exemplar-theoretic account of intonation, as discussed in section 10.4.2. Second, if a phonological perspective is preferred, the PaIntE model can be used to automatically label intonation events with an accuracy close to that of human labelers.

10.6 Conclusion

The PaIntE model can be used, in an analysis mode, to approximate the shapes of natural F0 curves and, in a synthesis mode, to generate F0 contours that sound convincingly like natural ones. The model considers the intonation structure of an utterance as consisting of a sequence of intonation events, which can be mapped to elements of the linguistic structure, with simple contour interpolations between these events.

By default, PaIntE considers ToBI categories as relevant intonation events. This is an obvious choice, given the prevalence of the autosegmental model in intonation research. ToBI categories are, at least, a good approximation of salient intonational events. Mapping PaIntE parameter values to these categories therefore facilitates the comparison of results across otherwise different approaches in phonological and phonetic intonation research. Moreover, evidence of a correlation between ToBI-like categories and the PaIntE parameters was found by Möhler (2001), who reported that the acoustic distance between natural and generated F0 contours is smaller when GToBI(S) categories are taken into account than when parameter values based on generic clustering were used. However, if no annotation of intonation events is available, PaIntE parameters can be extracted for each syllable in a given utterance. This approach was taken, for instance, in first-language acquisition studies with young children whose productions cannot be described adequately by means of adult ToBI-like categories (Lintfert and Möbius 2012).

In this chapter, we have presented the motivation behind the PaIntE modeling approach and its mathematical formulation. The PaIntE function comprises a small number of parameters with a linguistic interpretation whose values are estimated, learned, and generalized from speech databases. We explained the procedure of extracting the parameters from observed F0 contours and how to predict them, for instance, in the context of speech synthesis. Furthermore, we discussed the relation between the PaIntE intonation events and intonational categories posited by autosegmental approaches to intonation modeling. Finally, we presented several recent studies employing the PaIntE model. They show that PaIntE is a valuable contribution to intonation research beyond speech synthesis, which can serve to answer research questions both in the autosegmental tradition and in an exemplar-theoretic framework.

Notes

1. This holds under the assumption that establishing intonation categories works analogously to the segmental domain, where segmental categories are motivated by providing minimal pairs of words that differ only in the segmental category and have different meanings.

2. There is no closed-form expression for the true location of *b*, so the peak location must be approximated using numerical methods.

3. Also, the HH*L and H*M accents were not frequent enough in our data to reliably estimate their densities.

4. Thanks to Jörg Mayer for suggesting word finality as a possible explanation for the earlier peak.

5. https://librivox.org.

6. Note that research papers cited in this section are not always explicit about which version of the kappa coefficient (e.g., Cohen's kappa, Fleiss' kappa) they have used.

7. Perfect agreement is said to occur when $\kappa \geq 0.81$ (Landis and Koch 1977).

8. Comparing these results is straightforward because the data are similar: they all report results for the two-class problem, deriving the accent status from ToBI labels. All corpora consist of news-style read speech by professional speakers. However, it should be noted that the corpora are from two different languages with different ToBI systems.

9. We have only reported consistencies for pitch accents here; it should be noted that consistency for boundary tones is usually higher than for pitch accents, indicating that the boundary categories are easier to identify than pitch accents and in that respect are less problematic in our view.

References

Barbisch, M., G. Dogil, B. Möbius, B. Säuberlich, and A. Schweitzer. 2007. "Unit Selection Synthesis in the SmartWeb Project." In *Sixth ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, edited by P. Wagner, J. Abresch, S. Breuer, and W. Hess, 304–309. https://www.isca-speech.org/archive_open/ssw6/index.html.

Baumann, S., and M. Grice. 2006. "The Intonation of Accessibility." *Journal of Pragmatics* 38 (10):1636–1657.

Baumann, S., C. Röhr, and M. Grice, M. 2015. "Prosodische (de-)Kodierung des Informationsstatus." *Zeitschrift für Sprachwissenschaft* 34 (1):1–42.

Beckman, M. E. 1996. “The Parsing of Prosody.” *Language and Cognitive Processes* 11 (1/2): 17–67.

Bruce, G. 1977. *Swedish Word Accent in a Sentence Perspective*. Lund, Sweden: Gleerup.

Büring, D. 1997. *The Meaning of Topic and Focus: The 59th Street Bridge Accent*. New York: Routledge.

Calhoun, S., J. Carletta, J. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver. 2010. “The NXT-Format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue.” *Language Resources and Evaluation* 44 (4): 387–419.

Calhoun, S., and A. Schweitzer. 2012. “Can Intonation Contours Be Lexicalised? Implications for Discourse Meanings.” In *Prosody and Meaning*, edited by G. Elordieta Alcibar and P. Prieto, 271–327. Berlin: Mouton DeGruyter.

Chen, A., E. den Os, and J. de Ruiter. 2007. “Pitch Accent Type Matters for Online Processing of Information Status: Evidence from Natural and Synthetic Speech.” *Linguistic Review* 24:317–344.

Cooper, W. E., and J. M. Sorensen. 1981. *Fundamental Frequency in Sentence Production*. New York: Springer.

Dilley, L., and M. Brown. 2005. “The RaP Labeling System, v. 1.0.” Unpublished manuscript. Michigan State University. http://speechlab.cas.msu.edu/RaP/RaP_Labeling_Guide_v1.0.pdf.

Dogil, G., and A. Schweitzer. 2011. “Quantal Effects in the Temporal Alignment of Prosodic Events.” In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by W.-S. Lee and E. Zee, 595–598. City University of Hong Kong. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/.

Eckart, K., A. Riester, and K. Schweitzer. 2012. “A Discourse Information Radio News Database for Linguistic Analysis.” In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann, 65–75. Berlin: Springer.

Féry, C. 1993. *The Meaning of German Intonational Patterns*. Tübingen, Germany: Max Niemeyer Verlag.

Féry, C., and F. Kügler. 2008. “Pitch Accent Scaling on Given, New and Focused Constituents in German.” *Journal of Phonetics* 36 (4):680–703.

Godfrey, J., Holliman, E., and McDaniel, J. 1992. “Switchboard: Telephone Speech Corpus for Research and Development.” In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520. https://doi.org/10.1109/ICASSP.1992.225858.

Goedemans, R. 1998. *Weightless Segments: A Phonetic and Phonological Study Concerning the Metrical Irrelevance of Syllable Onsets*. Leiden, the Netherlands: Holland Academic Grahics.

Goldinger, S. D. 1996. “Words and Voices: Episodic Traces in Spoken Word Identification and Recognition Memory.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22:1166–1183.

Goldinger, S. D. 1997. “Words and Voices—Perception and Production in an Episodic Lexicon.” In *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix, 33–66. San Diego: Academic Press.

Goldinger, S. D. 1998. “Echoes of Echoes? An Episodic Theory of Lexical Access.” *Psychological Review* 105:251–279.

Goldsmith, J. A. 1976. “Autosegmental Phonology.” PhD diss., MIT.

Grabe, E. 1998. “Pitch Accent Realizations in English and German.” *Journal of Phonetics* 26:129–143.

Grice, M., and S. Baumann, S. 2002. “Deutsche Intonation und GToBI.” *Linguistische Berichte* 191:267–298.

Grice, M., S. Baumann, and R. Benzmüller. 2005. “German Intonation in Autosegmental-Metrical Phonology.” In *Prosodic Typology*, edited by S.-A. Jun, 53–83. Oxford: Oxford University Press.

Grice, M., M. Reyelt, R. Benzmüller, J. Mayer, and A. Batliner. 1996. “Consistency in Transcription and Labelling of German Intonation with GToBI.” In *Proceedings of the Fourth International Conference on Spoken Language Processing*, edited by H. T. Bunnell and W. Idsardi, 1716–1719. https://www.isca-speech.org/archive/icslp_1996/.

Gussenhoven, C. 1984. *On the Grammar and Semantics of Sentence Accents*. Dordrecht, the Netherlands: Foris Publications.

Gussenhoven, C. 2005. “Transcription of Dutch Intonation.” In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun, 118–145. Oxford: Oxford University Press.

Gussenhoven, C., and A. Rietveld. 1988. “Fundamental Frequency Declination in Dutch: Testing Three Hypotheses” *Journal of Phonetics* 16:355–369.

Gut, U., and P. S. Bayerl. 2004. “Measuring the Reliability of Manual Annotations of Speech Corpora.” In *Proceedings of Speech Prosody*, edited by B. Bel and I. Marlien, 565–568. https://www.isca-speech.org/archive/sp2004/.

Hasegawa-Johnson, M., K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, and T. Yoon. 2005. “Simultaneous Recognition of Words and Prosody in the Boston University Radio Speech Corpus.” *Speech Communication* 46 (3–4):418–439.

House, D. 1996. “Differential Perception of Tonal Contours through the Syllable.” In *Proceedings of the Fourth International Conference on Spoken Language Processing*, edited by H. T. Bunnell and W. Idsardi, 2048–2051. https://www.isca-speech.org/archive/icslp_1996/.

House, D. 1997. “Temporal Alignment Categories of Accent-Lending Rises and Falls.” In *Proceedings of Eurospeech*, edited by G. Kokkinakis, N. Fakotakis, and E. Dermatas, 879–882. https://www.isca-speech.org/archive/eurospeech_1997/.

Jilka, M., G. Möhler, and G. Dogil. 1999. “Rules for the Generation of ToBI-Based American English Intonation.” *Speech Communication* 28:83–108.

Johnson, K. 1997. “Speech Perception without Speaker Normalization: An Exemplar Model.” In *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullennix, 145–165. San Diego: Academic Press.

Kelly, N., and K. Schweitzer. 2015. “Examining Lexical Tonal Contrast in Norwegian Using Intonation Modelling.” In *Proceedings of the Eighteenth International Congress of Phonetic Sciences*. Glasgow: University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/proceedings.html.

Kelly, N. E., and R. Smiljanić. 2014. “The Effect of Focus on Norwegian Tonal Accent.” In *Proceedings of the Fourth International Symposium on Tonal Aspects of Languages*, edited by C. Gussenhoven, Y. Chen, and D. Dediu, 95–99. https://www.isca-speech.org/archive/tal_2014/.

Kohler, K. J. 1990. “Macro and Micro F0 in the Synthesis of Intonation.” In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. E. Beckman, 115–138. Cambridge: Cambridge University Press.

Kohler, K. 1991. “A Model of German Intonation.” In *Studies in German Intonation, Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel* 25:295–360.

Kollmeier, B., T. Brand, and B. T. Meyer. 2008. *Perception of Speech and Sound*. Berlin: Springer.

Kristoffersen, G. 2006. “Tonal Melodies and Tonal Alignment in East Norwegian.” In *Nordic Prosody IX*, edited by G. Bruce and M. Horne, 157–166, Frankfurt: Peter Lang.

Lacerda, F. 1995. “The Perceptual-Magnet Effect: An Emergent Consequence of Exemplar-Based Phonetic Memory.” *Proceedings of the Thirteenth International Congress of Phonetic Sciences,* edited by K. Elenius and P. Branderud. 2:140–147. Stockholm: KTH and Stockholm University.

Ladd, D. R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Ladd, D. R., I. Mennen, and A. Schepman. 2000. “Phonological Conditioning of Peak Alignment in Rising Pitch Accents in Dutch.” *Journal of the Acoustical Society of America* 107 (5): 2685–2696.

Landis, J. R., and G. G. Koch. 1977. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33 (1):159–174.

Lintfert, B., and B. Möbius. 2012. “Describing the Development of Intonational Categories Using a Target-Oriented Parametric Approach.” In *Proceedings of Interspeech 2012*, 1965–1968. https://www.isca-speech.org/archive/interspeech_2012/.

Mayer, J. 1995. *Transcription of German Intonation—the Stuttgart System*. Technical report. Stuttgart, Germany: Institute of Natural Language Processing, University of Stuttgart.

Mayo, C., M. Aylett, and D. R. Ladd. 1997. “Prosodic Transcription of Glasgow English: An Evaluation Study of GlaToBI.” In *Proceedings of an ESCA Workshop: Intonation: Theory, Models and Applications*, edited by G. Kouroupetroglou and G. Carayiannis, 231–234.

Möhler, G. 2001. “Improvements of the PaIntE Model for F0 Parametrization.” Unpublished manuscript. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.6756.

Möhler, G., and A. Conkie. 1998. “Parametric Modeling of Intonation Using Vector Quantization.” In *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, 311–316. https://isca-speech.org/archive_open/ssw3/.

Ostendorf, M., P. Price, and S. Shattuck-Hufnagel. 1996. *Boston University Radio Speech Corpus*. Philadelphia: Linguistic Data Consortium.

Peters, J. 2014. *Intonation*. Heidelberg: Winter.

Pierrehumbert, J. 1979. “The Perception of Fundamental Frequency Declination.” *Journal of the Acoustical Society of America* 66:363–369.

Pierrehumbert, J. 1980. “The Phonology and Phonetics of English Intonation.” PhD diss., MIT.

Pierrehumbert, J. 1981. “Synthesizing Intonation.” *Journal of the Acoustical Society of America* 70:985–995.

Pierrehumbert, J. 2001. “Exemplar Dynamics: Word Frequency, Lenition and Contrast.” In *Frequency and the Emergence of Linguistic Structure*, edited by J. Bybee and P. Hopper, 137–157. Amsterdam: Benjamins.

Pierrehumbert, J. 2003. “Probabilistic Phonology: Discrimination and Robustness.” In *Probability Theory in Linguistics*, edited by R. Bod, J. Hay, and S. Jannedy, 177–228. Cambridge, MA: MIT Press.

Pierrehumbert, J., and J. Hirschberg. 1990. “The Meaning of Intonational Contours in the Interpretation of Discourse.” In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Pitrelli, J., M. Beckman, and J. Hirschberg. 1994. “Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework.” In *Proceedings of the Third International Conference on Spoken Language Processing*, 123–126. https://www.isca-speech.org/archive/icslp_1994/.

Rapp, S. 1998. “Automatisierte Erstellung von Korpora für die Prosodieforschung.” PhD diss., University of Stuttgart.

Rosenberg, A. 2009. “Automatic Detection and Classification of Prosodic Events.” PhD diss., Columbia University.

Schweitzer, A. 2011. “Production and Perception of Prosodic Events—Evidence from Corpus-based Experiments.” PhD diss., Universität Stuttgart.

Schweitzer, A., and B. Möbius. 2009. “Experiments on Automatic Prosodic Labeling.” In *Proceedings of the Tenth Annual Conference of the International Speech Communication Association*, 2515–2518. https://www.isca-speech.org/archive/interspeech_2009/.

Schweitzer, K., M. Walsh, S. Calhoun, H. Schütze, B. Möbius, A. Schweitzer, and G. Dogil. 2015. “Exploring the Relationship between Intonation and the Lexicon: Evidence for Lexicalised Storage of Intonation.” *Speech Communication* 66:65–81.

Sproat, R., ed. 1998. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Dordrecht, the Netherlands: Kluwer.

Sridhar, V. K. R., S. Bangalore, and S. Narayanan. 2008. “Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework.” *IEEE Transactions on Audio, Speech, and Language Processing* 16 (4): 797–811. https://doi.org/10.1109/TASL.2008.917071.

Syrdal, A. K., J. Hirschberg, J. McGory, and M. Beckman. 2001. “Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody.” *Speech Communication* 33:135–151.

Syrdal, A. K., and J. McGory. 2000. “Inter-Transcriber Reliability of ToBI Prosodic Labeling.” *Proceedings of the Sixth International Conference on Spoken Language Processing* 3:235–238. https://www.isca-speech.org/archive/icslp_2000/.

Syrdal, A. K., C. W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K.-S. Lee, and M. J. Makashay. 2000. “Corpus-Based Techniques in the AT&T NextGen Synthesis System.” *Proceedings of the Sixth International Conference on Spoken Language Processing* 3:410–413. https://www.isca-speech.org/archive/icslp_2000/.

Taylor, P. 1998. “The Tilt Intonation Model.” In *Proceedings of the Fifth International Conference on Spoken Language Processing*, 1383–1386. https://www.isca-speech.org/archive/icslp_1998/.

Taylor, P. 2009. *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

Taylor, P., R. Caley, A. W. Black, and S. King. 1999. *Edinburgh Speech Tools Library*. System Documentation Edition 1.2, for 1.2.0. June 15. http://festvox.org/docs/speech_tools-1.2.0/.

Terken, J., and J. Hirschberg. 1994. “Deaccentuation of Words Representing ‘Given’ Information: Effects of Persistence of Grammatical Role and Surface Position.” *Language and Speech* 37:125–145.

van Santen, J. P. H., and B. Möbius. 2000. “A Quantitative Model of F0 Generation and Alignment.” In *Intonation: Analysis, Modelling and Technology*, edited by A. Botinis, 269–288. Dordrecht, the Netherlands: Kluwer.

Wade, T., G. Dogil, H. Schütze, M. Walsh, and B. Möbius. 2010. “Syllable Frequency Effects in a Context-Sensitive Segment Production Model.” *Journal of Phonetics* 38 (2): 227–239.

Wahlster, W. 2004. “SmartWeb: Mobile Applications of the Semantic Web.” In *KI 2004: Advances in Artificial Intelligence*, edited by S. Biundo, T. Frühwirth, and G. Palm, 50–51. Berlin: Springer.

Walsh, M., B. Möbius, T. Wade, and H. Schütze. 2010. “Multilevel Exemplar Theory.” *Cognitive Science* 34:537–582.

Witten, I. H., and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann.

Yoon, T.-J., S. Chavarría, J. Cole, and M. Hasegawa-Johnson. 2004. “Intertranscriber Reliability of Prosodic Labeling on Telephone Conversation Using ToBI.” In *Proceedings of the Eighth International Conference on Spoken Language Processing*, 2729–2732. https://www.isca-speech.org/archive/interspeech_2004/.

11 The PENTA Model: Concepts, Use, and Implications

Yi Xu, Santitham Prom-on, and Fang Liu

11.1 Introduction

Speech is a communication system for transmitting information from one human being to another.[1] The information transmitted is rich and multifaceted, but it is coded by an articulatory system in such a way that the listener can readily decode it. These facts, which may seem too obvious to be worth restating, are the premise of the articulatory-functional view of speech that forms the basis of the parallel encoding and target approximation (PENTA) model (Xu 2005). PENTA is therefore a theory of how multiple layers of information are effectively conveyed through prosody with a neurally controlled biomechanical system. In other words, PENTA is about how prosody works as a communication system, how it can be learned, and how it goes through changes over time—in short, how it operates. The mission of PENTA therefore differs from those of many other theories that focus on directly accounting for observed prosodic forms. By focusing on operation as its primary goal, PENTA accounts for prosodic forms only as a by-product, rather than as an end in itself.

11.2 The Conceptual Framework

Beyond the basic facts just stated, PENTA makes a number of assumptions that are highly hypothetical. The first is that syllable-synchronized sequential target approximation (the TA part of the model) is the rudimentary mechanism of speech prosody, based on which all of the information coding is done. The second is that prosody conveys multiple layers of information simultaneously, through encoding schemes that are in parallel to each other, that is, without a hierarchical structure (the PE part of the model). Third, the phonetics of the encoding schemes are specified parametrically rather than based on symbolic representations. Due to their hypothetical nature, each of these assumptions needs independent justifications, which we will provide after a brief sketch of the model.

Figure 11.1 is a schematic of PENTA in its most general form, representing not only prosody but also other aspects of speech (Xu and Liu 2012). The first block from the left represents communicative functions that are conveyed by speech. The functions are parallel to each other, as illustrated by the nonhierarchical stacking in the schematic. The second block represents encoding schemes that are associated with the communicative functions. The schematization here makes it clear that communicative functions do not control surface acoustics directly but through a set of specific encoding schemes. The encoding schemes can be highly stylized and language-specific or more gradient and universal. The third block represents the articulatory parameters that are

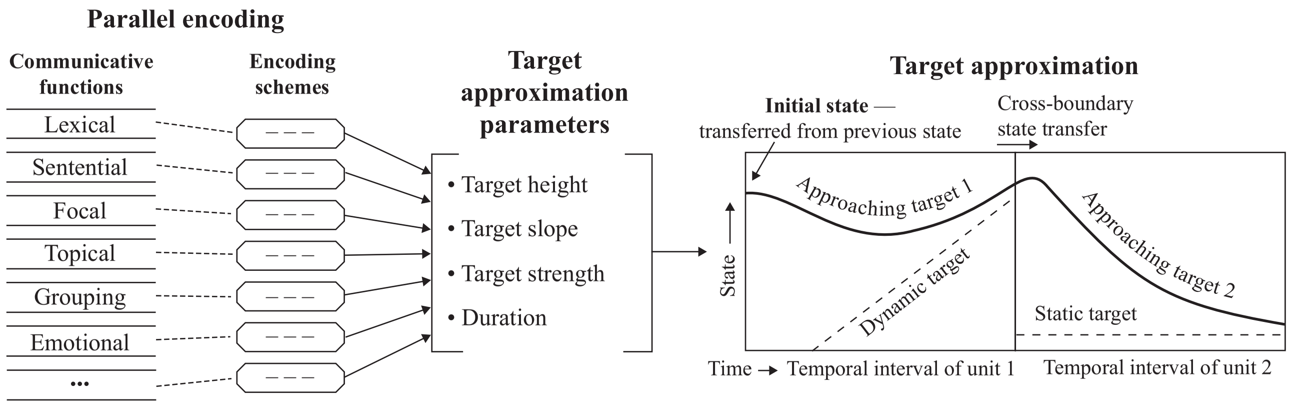

Parallel encoding
Communicative functions
Encoding schemes
Lexical
Sentential
Focal
Topical
Grouping
Emotional
...
Target approximation parameters
• Target height
• Target slope
• Target strength
• Duration
Target approximation
Initial state — transferred from previous state
Cross-boundary state transfer
Approaching target 1
Approaching target 2
Dynamic target
Static target
State
Time → Temporal interval of unit 1
Temporal interval of unit 2

Figure 11.1
A schematic sketch of the PENTA model (Xu 2005; Xu and Liu 2012; Xu and Wang 2001).

specified by the encoding schemes. These parameters control the articulatory process of target approximation represented by the fourth block. This biomechanical process ultimately generates surface acoustics.

The TA model, as depicted in the rightmost block in figure 11.1, assumes that each syllable is assigned an underlying pitch target specified in terms of not only height but also slope. The surface contour is then the result of sequential approximations of successive targets, each articulated in synchrony with a syllable. At the boundary between adjacent syllables, the final articulatory state of the earlier syllable is transferred to the next syllable. Such transfer often results in a delay of the surface alignment of a turning point. The model therefore has no specifications for the temporal alignment of surface turning points, unlike in some other models (e.g., Hirst 2005; Pierrehumbert 1980; Taylor 2000).

11.2.1 Articulatory Mechanisms

Like the theory as a whole, the articulatory aspect of PENTA also starts from self-evident facts. One of the most basic is that, given the need to use different articulatory states to represent different prosodic components (because information has to be coded by differential representation), transitional movements between the states are inevitable, and each movement takes time. Two empirical questions need to be answered about these movements:

Question 1: How much time does each take? If the amount of time is largely negligible, then there is little need to take it seriously in theoretical considerations.
Question 2: What is the manner of the transitional movements? Knowledge about this may help to explain many details in the observed surface acoustics.

The second basic fact is that the laryngeal movements that generate F_0 contours have to co-occur with supralaryngeal movements that generate segments, and the two movements have to be coordinated in time. For this, there are two empirical questions:

Question 3: What is the basic mechanism of the temporal coordination?
Question 4: How much freedom does the speaker have in this kind of coordination?

With regard to question 1 (how much time the F_0 movements have to take), the empirical findings (Sundberg 1979; Xu and Sun 2002) have shown that the time needed for F_0 movements due to change of laryngeal state is not negligible. The following quasi-linear relations are found between the size of F_0 movement and the mean minimum time it takes to complete the movement:

$$\textit{Rise:}\ \ t_r = 89.6 + 8.7\ d \qquad (11.1)$$

$$\textit{Fall:}\ \ t_f = 100.04 + 5.8\ d \qquad (11.2)$$

where d is the F_0 excursion size in semitones, and t is the duration in milliseconds. These equations show that it takes about one hundred milliseconds to make even the smallest pitch movement, rising or falling. Given that the average speech rate is about five to seven syllables per second, meaning that each syllable takes about 143 to 200 milliseconds, a significant portion of each syllable has to be used for pitch transitions.

Regarding the manner of the transitional movements (question 2), systematic examination of lexical tones in Mandarin produced in connected speech has provided relevant clues. As can be seen in figure 11.2, when a tone is preceded by different tones, the corresponding F_0 contours all asymptotically approach a linear trajectory that is characteristic of its underlying properties: high level for the high tone, rising for the rising tone, and falling for the falling tone. Cross-tonal transitional movements can therefore be characterized as asymptotic approximation of the underlying target (Xu and Wang 2001).

As for the basic mechanism of the temporal coordination of laryngeal and supralaryngeal movements (question 3), figure 11.2 also provides relevant clues. We can see that the divergence between different tones is the largest at the beginning of the syllable and smallest at the end of the syllable. This indicates that the movement toward a tonal target starts from the onset of a syllable and ends at its offset. Thus, the target approximation movement of a tone seems to be synchronized with respect to the syllable. In other words, each target approximation movement occurs strictly within the syllable that carries the tone. These observations have led to PENTA's core assumption about the basic articulatory mechanism of F_0 production: syllable-synchronized sequential target approximation (Xu 2005).

There are actually two aspects to this basic articulatory mechanism: pitch target in every syllable and syllable-synchronized sequential target approximation. "Target in every syllable" means that even if there appears to be a global contour over more than one syllable, each syllable still needs to be assigned a pitch target. This is based on the articulatory consideration that it is physically impossible, as illustrated in figure 11.3, to first generate a pitchless syllable (lower left panel) and a carrierless F_0 contour (upper left panel), and then combine them to form the full surface acoustic signal (right panels). It is imaginable that the F_0 contours in the upper left panel and the formant trajectories in the lower left panel could be first formed in the brain, and then the articulation process faithfully reproduces them during articulation. In fact, as already mentioned, much of the F_0 movements result from articulatory transitions between the ideal pitch targets, and their slow speed (relative to the syllable) is due to the limit of maximum speed of pitch change. The same is shown to be true of formant movements as well (Cheng and Xu 2013). Thus, the F_0 and formant contours in figure 11.3 are mostly a by-product of physical inertia. Therefore, if the transitions were already represented in the brain-generated F_0 and formant commands sent to the muscles, the effect of inertia would be applied twice!

Thus, it is unlikely that continuous surface F_0 contours are generated independent of the segmental events and then added to the segmental string during articulation. Instead, it is more likely that, at the control level, each syllable is specified with all the underlying articulatory goals associated with it, including segmental targets, pitch targets, and even phonation (voice quality) targets. This is illustrated in the left block of figure 11.4 for pitch and formants. Here the formant patterns are representations of the corresponding vocal tract shapes that are presumably the actual targets. The articulation process then implements all the targets in tandem, all through target approximation (top right). This biomechanical process ultimately generates continuous surface F_0 and formant trajectories (bottom), which consist of mostly transitions toward the respective targets. Thus, every syllable, before its articulation, would have to be assigned both segmental and suprasegmental targets as control signals for the articulatory system. And, importantly, the effects of inertia are applied only once: during articulatory execution, the final stage in the production chain.

Further support for pitch target for every syllable comes from the finding that not only stressed syllables, but also unstressed syllables in English and the neutral tone

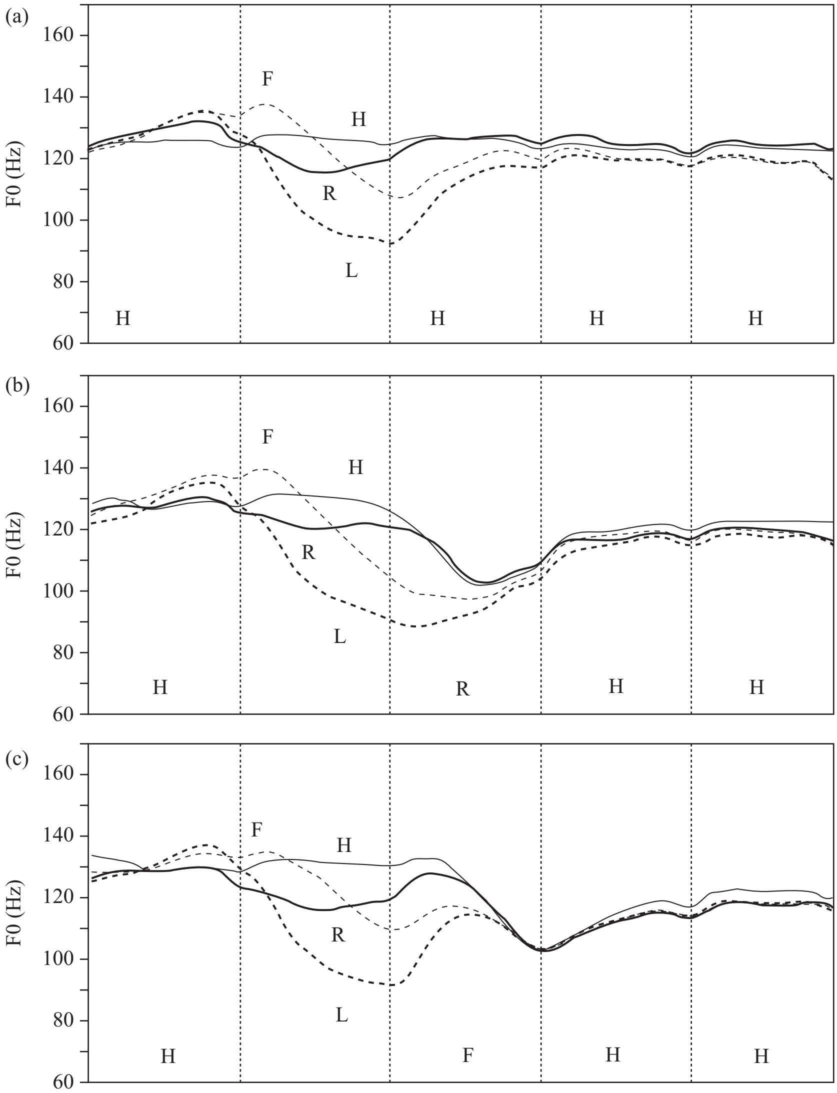


Figure 11.2

Mandarin tones produced in various tonal contexts. These are mean time-normalized F_0 contours of 猫咪/迷/米/蜜 摸/拿/卖 猫咪 (Kitty/Cat-fan/Cat rice/Cat honey strokes/picks up/sells kitty). Each contour is an average of forty tokens said by four male speakers of Beijing Mandarin (five repetitions by each). In all three plots, vertical lines indicate syllable boundaries. The tone names H, R, L, and F stand for high, rising, low, and falling tones, respectively. *Source*: Adapted from Xu (1999).

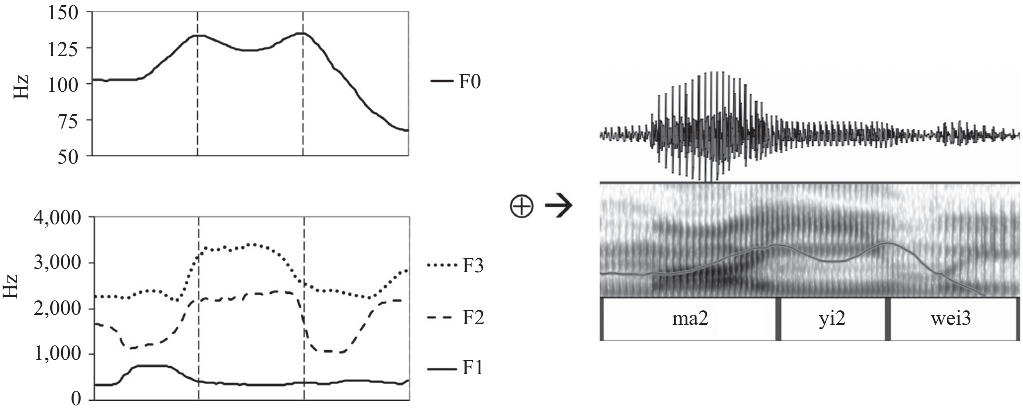


Figure 11.3
(Left) Continuous F_0 (top) and formant (bottom) tracks of the Mandarin utterance *(bi3) ma2 yi2 wei3 (shan4)* (More hypocritical than Aunt Ma). (Right) Waveform, spectrogram, and F_0 track of the same utterance. *Source*: Raw data courtesy of Xu (2007).

in Mandarin, show signs of functionally contrastive pitch targets. For example, when a word was focused in English, the unstressed syllables following a stressed syllable actively lowered pitch, as if they were part of postfocus domain. But the lowering is not as fast as in a postfocus stressed syllable (Xu and Xu 2005). This indicates that unstressed syllables are assigned postfocus targets, but with a weak target approximation strength. As found in both acoustic analysis (Chen and Xu 2006; Xu and Xu 2005) and computational modeling (Liu et al. 2013; Xu and Prom-on 2014), such weak strength can account for the high variability (and hence a seeming lack of target) of the pitch of the unstressed syllables in English and the neutral tone in Mandarin.

The other hypothesized aspect of the pitch production mechanism is the syllabic synchronization of target approximation movements. That is, laryngeal and supralaryngeal movements are synchronized with respect to the syllable (Xu 2020; Xu and Liu 2006, 2012). This assumption is motivated not only by observations like those shown in figure 11.2, but also by findings that the motor system is able to coordinate multiple movements at a fast speed only through full synchrony (Kelso 1984; Kelso, Southard, and Goodman 1979; Mechsner et al. 2001). This synchrony constraint could be due to a general problem in motor control, that is, the high dimensionality of the motor system makes the control of any motor action extremely challenging (Bernstein 1967; Latash 2012). Bernstein (1967) proposes that this problem can be alleviated by functionally freezing degrees of freedom (DOF) during motor learning. The freezing of DOF is analogous to allowing the wheels of a car to be controlled by only a single steering wheel. This makes the movements of the wheels fully synchronized and their DOFs merged. In the case of speech, it is possible that the syllable is an evolved motor synchronization mechanism to solve the problem of multigestural encoding of information. As conceptualized in the synchronization model of the syllable (Xu 2020; Xu and Liu 2006, 2012), all the articulatory target approximation movements are synchronized with respect to the syllable, except that syllable-initial consonants complete their movements earlier than both the vocalic and laryngeal movements.

There are a few issues that the current version of the TA model, especially its quantitative implementation via qTA (Prom-on, Xu, and Thipakorn 2009), has not yet fully

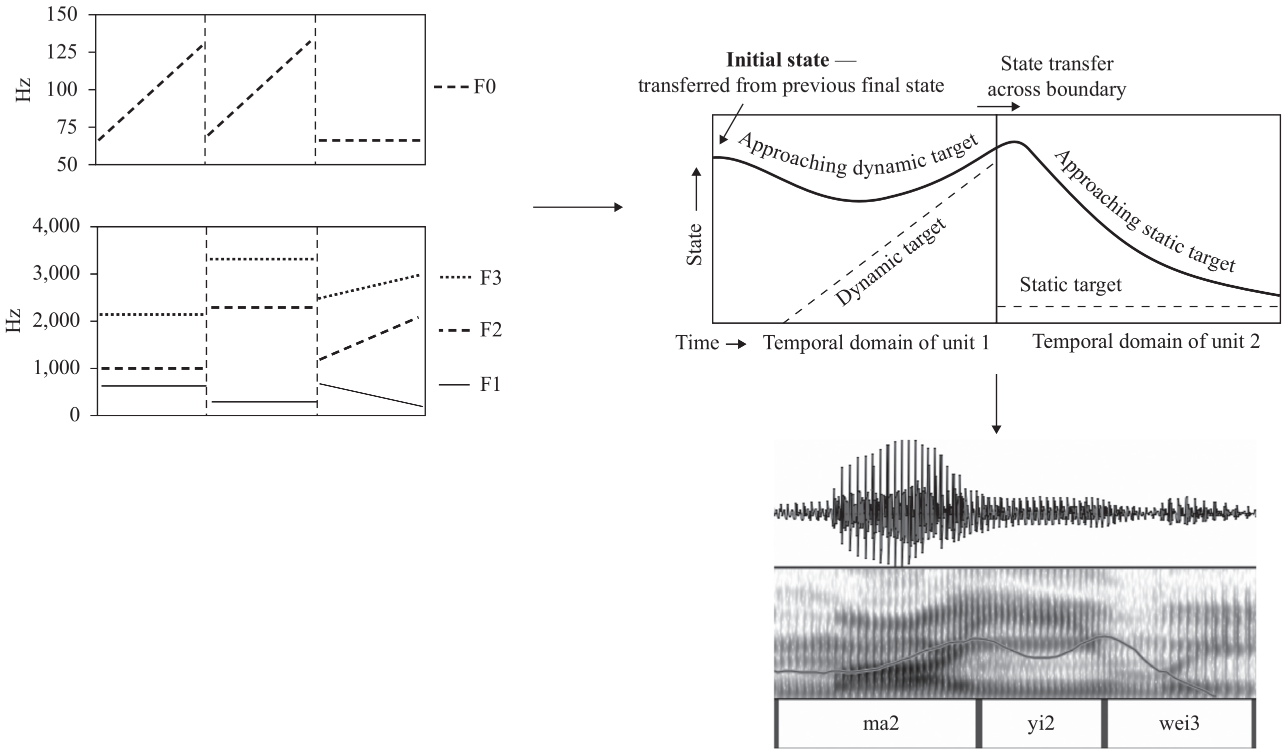

150
125
100
75
50
Hz
F0
4,000
3,000
2,000
1,000
0
Hz
F3
F2
F1
Initial state —
transferred from previous final state
State transfer
across boundary
Approaching dynamic target
Approaching static target
Dynamic target
Static target
State
Time
Temporal domain of unit 1
Temporal domain of unit 2
ma2
yi2
wei3

Figure 11.4
(Left) Hypothetical underlying pitch (top) and formant (bottom) targets for the Mandarin utterance shown in figure 11.3. (Right) The TA model (Xu and Wang 2001). (Bottom) Waveform, spectrogram, and F_0 track of the same utterance that are presumably generated with the underlying targets through an articulatory process similar to the TA model. *Source*: Raw data courtesy of Xu (2007).

addressed. The first is the delay in target approximation when a target is dynamic. It has been found that in Mandarin, the final slope of a rising or falling tone remains consistent when syllable duration is excessively long (Xu 1998, 2001). A similar finding has been reported for the final slope of formant trajectories in English diphthongs (Gay 1968). A possible articulatory mechanism is that speakers have the ability to adjust their articulatory effort over time to achieve consistent movement velocity by the end of a syllable. But this mechanism has not yet been implemented in the current version of qTA.

The second issue is the possibility of having two consecutive pitch targets within one syllable. An example is that the citation form of the low tone (tone 3) in Mandarin has a rising movement following the initial low-approaching movement. Although the entire shape looks like that of the rising tone, which also has an initial dip followed by a rise, the rising tail in the low tone is entirely missing when the tone is followed by any other tone. Thus, the final tail of the low tone is optional, and the underlying target is unlikely to be similar to that of the rising tone. Rather, it is possible that the citation form of the low tone consists of a low target followed by either a mid/high or rising target. Such consecutive targets in a syllable would still obey the synchronization constraint because the onset of the first target and the offset of the second one would coincide with the syllable onset and offset, respectively. Like delayed target approximation, the mechanism of consecutive targets also has yet to be implemented in qTA.

The third issue is the phenomenon of postlow bouncing. That is, after a tone with a very low pitch, the F_0 of the following syllables sometimes "bounces up" before returning to the same level as other tones. The bouncing is most readily observed if the postlow tone is weak, as when it is a neutral tone (Chen and Xu 2006). When the following tone is not weak, the effect is observable only when the low-pitched tone is focused. We have speculated that such bouncing is due to a temporary loss of antagonistic muscle balance when the extrinsic laryngeal muscles, which are engaged mainly in pitch lowering (sternohyoids, sternothyroids, thyrohyoids), suddenly stop contracting, resulting in an abrupt increase of vocal fold tension since the cricothyroids, the only muscles that lengthen the vocal folds (Zemlin 1988), are in contraction. We have been able to simulate postlow bouncing by increasing the amount of acceleration (second derivative of F_0) in the transferred state at the junction between the low tone and the following tone. The simulation worked well (Prom-on, Liu, and Xu 2012) but has not yet been included in modeling tools such as qTAtrainer and PENTAtrainer.

Closely related to postlow bouncing is the fourth issue: prelow raising. This is a phenomenon (also known as anticipatory raising, anticipatory dissimilation or H raising) whereby a tone or a pitch accent with a low pitch component raises the pitch of the preceding syllable (Connell and Ladd 1990; Hyman 1993; Gandour, Potisuk, and Dechongkit 1994; Laniran 1992; Laniran and Gerfen 1997; Lee, Xu, and Prom-on 2017; Xu 1997). Given that it is found in a number of unrelated languages, the phenomenon is likely due to a universal articulatory mechanism, although its exact nature is not yet clear. Our current theory is that it is a preparation for a sharp F_0 lowering movement, which is similar to drawing back the arm in preparation for throwing an object over a long distance. Prelow raising is not explicitly incorporated in qTAtrainer or PENTAtrainer either. However, its effect can be partially simulated by allowing tonal pitch targets to be conditioned by the upcoming tone in computational modeling.

Finally, it is well known that F_0 is temporally perturbed (mostly upward) at the voice onset after an obstruent consonant (Hombert 1977; Silverman 1986). The perturbation consists of a very brief (lasting about thirty milliseconds) transient effect (Xu and Wallace 2004; Xu and Xu 2003) and an optional longer-lasting vocal tension effect

(Stevens 1998; Xu and Wallace 2004). Again, these mechanisms have not yet been incorporated into our computational tools.

However, both qTAtrainer and PENTAtrainer do allow (and actually recommend) users to set the entire syllable, including any voiceless consonant, as the domain of target approximation. This strategy has generated F_0 contours that better match those of the original than when target approximation domain is limited to the rime only (Xu and Prom-on 2014).

11.2.2 Communicative Functions

Again following the principle of starting from basic facts, we first recognize that it is indisputable that many meanings are conveyed through prosody. But questions are open as to what exactly those meanings are and how each of them is prosodically coded. One of the key issues is whether meanings are coded directly, or through the intermediary of a prosodic phonology. For this there is a need to go at least as far back as the Saussurean view of linguistic sign (de Saussure 1916). The Saussurean view emphasizes the distinction between meanings and their bearers, which are meaningless, that is, between signifier and signified. Linguistic units in speech are therefore signified-signifier or function-form unities. Although this view is already deeply entrenched in many aspects of modern linguistic theories, the difficulty of prosody research suggests that further theoretical clarifications are needed. The first is that, of the two sides of the signified-signifier coin, is the signified—the communicative function, or the signifier—the phonetic form, that should provide the primary definition of a prosodic category? This point may not appear critical in the case of lexical items, because the function-form unities are relatively clear, given the easy delineation of lexical items. In prosody, however, the function-form unities are almost never clearly delineated. In this case, should function or form be given primacy when defining prosodic units? PENTA, similar to some other function-oriented models (e.g., Bailly and Holm 2005), holds that, when in doubt, it should be function rather than form that is the primary guide to the delineation of prosodic units.

A further complication of prosody is the multiplicity of meanings it conveys, which the classic Saussurean view again cannot easily address. For example, in the case of tone languages like Mandarin or Swedish or a pitch accent language like Japanese, pitch is used to form tones or pitch accents, which serve to distinguish words that are segmentally identical. But pitch is also used to encode focus that highlights a particular component of an utterance (Bolinger 1972), and sentence type that indicates whether the speaker is making a statement or asking a question (Eady and Cooper 1986). It is therefore not easy to identify through direct F_0 observations, as typically done in form-first approaches (e.g., Pierrehumbert and Hirschberg 1990), which parts are the signifiers of the three meanings, respectively.

Based on our recognition of these inherent difficulties, the development of PENTA has followed a two-pronged strategy. First, we have persistently tried to establish function-form unities one at a time by starting from a reasonably clear functional definition of each category and then empirically discover (or rule out) the prosodic means used by each language to encode it. As will be seen next, the prosodic means are not limited to temporally separated units as in the case of lexical items. Rather, they are often encoded by modifying existing forms that are already specified by other functions. For example, the basic shapes of a pitch contours of individual syllables are already specified by lexical tones in a tone language. So focus and sentence type can only be encoded by modifying these basic shapes in one way or another.

Second, because of the obscurity of the function-form relation in prosody, we have always tried to empirically discover, rather than either stipulate for theory-internal reasons or conceptually assume, the exact means with which each communicative function is encoded in prosody. For each function, reliable methods of triggering its occurrence and controlling its categorical contrasts are established, and experiments are conducted to uncover the phonetic realization of the functional contrasts, following all the necessary precautions needed for proper empirical studies, including, in particular, the use of minimal pairs and attention to details (Xu 2011b).

This has been facilitated by a method that has allowed us to systematically compare continuous F_0 contours in fine detail (Xu 2011b, 2013). The key component of the method is the extraction of time-normalized continuous F_0 contours. Time normalization takes the same number of F_0 measurements from each equivalent temporal interval, for example, the syllable, regardless of the interval duration. The time-normalized contours can then be averaged across repetitions and even across speakers. Such averaging removes variations from individual utterances that may mask the common characteristics that we look for. For example, the target approximation characteristic of tonal realization is brought to a clear view by the averaged F_0 contours in figure 11.2 because the contextual variants of the same tone with complete trajectories are pitted against each other in the same graph.[2]

These strategies have allowed us to see how it is possible for F_0 to encode multiple communicative functions simultaneously. As can be seen in figures 11.5 and 11.6, in both Mandarin and English, focus and sentence type can be realized on top of the lexical functions of tone and word stress by modifying the height and shape of local F_0 contours. These sentences were spoken with sentence-initial (in figure 11.5 only), sentence-medial, or sentence-final focus and as either statements or questions. As can be seen, the realization of focus and modality exhibits F_0 patterns that are best described in terms of their interactions both with each other and with lexical stress:

1. Focus is characterized by a robust postfocus pitch-range shift, with the direction of the shift dependent on modality as well as language. It is downward in statements in both languages. In questions, the shift is upward in English but downward in Mandarin. In the latter case, however, the downward shift is less sharp in questions than in statements, thus allowing Mandarin to still distinguish statements from questions under focus.
2. In English, focus and modality interact with lexical stress to determine the microproperties of syllabic pitch targets. For on-focus word-final stressed syllables, the target slope is falling in statements but rising in questions (*job* in figure 11.9b and d). For on-focus, nonfinal stressed syllables, the target slope is rising in both statements and questions.
3. In Mandarin, the contribution of modality is most clearly seen from the location of focus, except in the sentence-final word, where the modality difference can be seen even when the sentence is focus-neutral.

The principle of empirical guidance in establishing the encoding schemes has also allowed us to see that not all prosodic functions have to be chiefly encoded by F_0. The function of boundary marking (or grouping, phrasing), for example, is primarily marked by timing properties, which include both syllable and pause duration. In general, a longer preboundary duration signifies greater separation of two adjacent groups (Wagner 2005; Xu and Wang 2009). Toward the end of a sentence, word duration

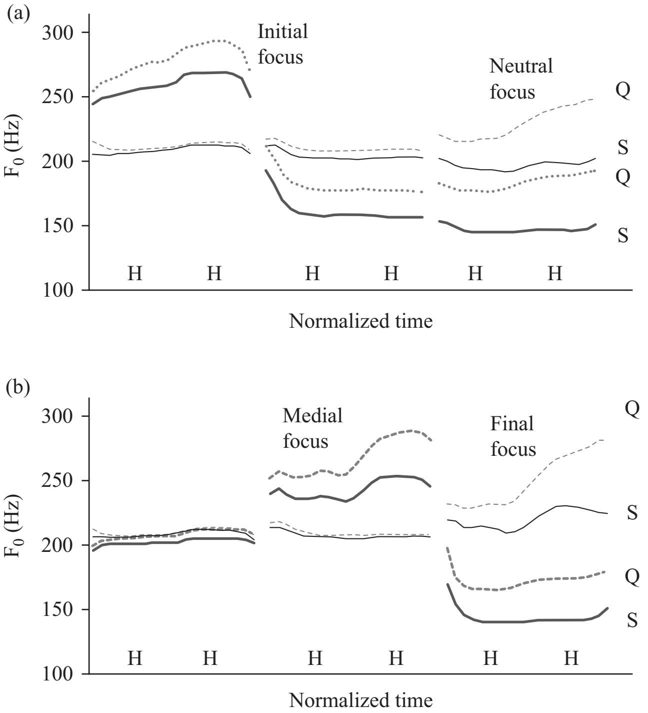


Figure 11.5

Mean F_0 contours of Mandarin sentence “Zhāng Wēi dānxīn Xiāo Yīng kāichē fāyūn” (张威担心肖英开车发晕; Zhang Wei is concerned that Xiao Ying may get dizzy when driving) spoken as either a statement or a question. (a) Either focus is on the sentence-initial word (thick lines) or there is no narrow focus (thin lines). (b) Focus is either sentence medial (thick lines) or sentence final (thin lines). The solid lines represent statements, and the dashed lines represent questions. S, statement; Q, question. *Source*: Data courtesy of Liu and Xu (2005).

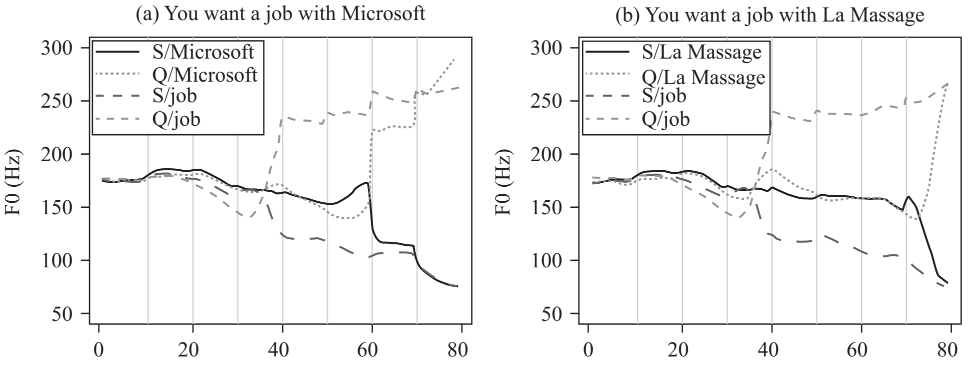


Figure 11.6

Mean F_0 contours of statements and questions in American English. The word after the virgule / is focused. *Source*: Data courtesy of Liu et al. (2013).

increases exponentially (Yuan, Liberman, and Cieri 2006). It is further hypothesized that syllable duration is combined with duration of silent pauses to serve as an affinity index that iconically encodes relational distance with temporal distance (Wang, Xu, and Zhang 2019; Xu 2009). It is also found, through a series of experimental studies, that emotion, attitude, and vocal attractiveness are encoded with a combination of pitch, spectral density, voice quality, and speech rate properties, based on a set of hypothetical bioinformational dimensions, and that such encoding is again done in parallel with other, more linguistic functions (Chuenwattanapranithi et al. 2008; Hsu and Xu 2014; Liu and Xu 2014; Noble and Xu 2011; Xu, Kelly, and Smillie 2013; Xu et al. 2013).

There are also various other communicative functions that may be prosodically encoded. But as reviewed in Xu (2011a), there may be more conceptually plausible communicative functions than there are consistent encoding schemes. The establishment of the latter, therefore, can only be done one by one through empirical studies.

11.3 PENTA as a Research Tool

As a theory of prosody, PENTA can be used in practice as a framework under which empirical studies can be conducted. Such usage may involve setting up hypotheses to be tested, selecting experimental factors to be manipulated, and interpreting data and results through systematic analyses. But PENTA can be also used as a research tool in its own right, thanks to its quantification into a computational model (Prom-on, Xu, and Thipakorn 2009). The current computational version is qTA, which is a third-order critically damped linear system, as shown in the following formula:

$$f_0(t) = (mt + b) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t} \tag{11.3}$$

Here, $f_0(t)$ is the surface F_0 of a syllable as a function of time. $mt+b$, the first term, is the underlying pitch target as a linear function of time t, with m representing the slope and b representing the height of the target. The second term represents the natural response of the system, in which the transient coefficients c_1, c_2, and c_3 are calculated based on the initial F_0 dynamic state and pitch target of the current syllable. Parameter λ represents the strength of the F_0 movement toward the target.

qTA also explicitly represents the state transfer between adjacent syllables by taking the final F_0 state of the preceding syllable in terms of its final F_0 $f_0(0)$, velocity $f_0'(0)$, and acceleration $f_0''(0)$ as the initial F_0 dynamic state of the current syllable. With this initial state, the three transient coefficients are computed with the following formulae:

$$c_1 = f_0(0) - b \; c_1 = f_0(0) - b \tag{11.4}$$

$$c_2 = f_0'(0) + c_1\lambda - m \tag{11.5}$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2 \tag{11.6}$$

A number of principles are behind the development of qTA. The first is to use as few free parameters as possible, and every free parameter should be communicatively meaningful, that is, usable by one or more encoding schemes. The second principle is that all the critical components of the model should be quantitatively implemented, so as to faithfully reflect the theoretical framework. Particularly worth mentioning is the specification of target slope *m*, which, as far as we know, is used only by qTA. The need for this parameter is based on findings that the final slope is one of the most consistent properties of a dynamic tone, as mentioned in 11.2.1. The third principle is that the model parameters should be learnable from real speech data, so that there are no missing links between the theoretical framework and the reality of continuous speech. This would also potentially allow the simulation of the acquisition of tone and intonation.

Note that qTA computes F_0 contours of only a single sequence of syllables, which is different from superpositional models (Fujisaki 1983; Bailly and Holm 2005) that compute multiple layers of F_0 contours and then combine them. There are no mathematical restrictions, however, on using units other than the syllable. But in our own practice, we have always used the syllable (Xu and Prom-on 2014) as the domain of local pitch targets. This scheme also means that for each syllable, there is only a single target that needs to be estimated (but see discussion in section 11.2.1 of the possibility of having consecutive targets within a single syllable), which has turned out to be a critical property for our understanding of the operation of speech prosody, as will be discussed later.

11.3.1 Computational Modeling Tools

Since the proposal of qTA, three computational tools have been developed to enable its conceptual exploration and quantitative testing. qTA_demo (www.homepages.ucl.ac.uk/~uclyyix/qTA/) is a web-based interactive programs that demonstrates how qTA works. Its interactive features make it a convenient tool for a quick impromptu test of an idea or a prediction based on the TA model.

The other two tools, qTAtrainer (www.homepages.ucl.ac.uk/~uclyyix/qTAtrainer/) and PENTAtrainer (www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer/), both use machine learning algorithms to automatically extract target parameters from real speech data through analysis-by-synthesis. In the analysis-by-synthesis paradigm, candidate targets are iteratively tested by applying them in the qTA function to generate continuous F_0 contours. The goodness of fit between the synthetic and natural contours is then used as the criterion in the selection of the targets (Prom-on, Xu, and Thipakorn 2009; Prom-on and Xu 2012). The quality of the F_0 generation is assessed by three means: (i) root mean squared error, which measures the discrepancy of the synthetic contours from the original contours in terms of point-by-point height difference; (ii) Pearson's *r*, which assesses how closely the overall shape of the synthetic contours correlates with that of the natural contours; and (iii) perceptual evaluation in terms of category identification (e.g., tone, focus) and naturalness.

Both qTAtrainer and PENTAtrainer allow predictive synthesis of F_0 contours using categorical parameters learned from training. They differ only in terms of how function-specific targets are obtained. qTAtrainer takes a two-phase approach. In phase 1, an optimal target is obtained for each syllable of each utterance by comparing the performance of all possible combinations of the three target parameters (b, m, and λ in equation (11.3)). The parameter set that achieves the best fit to the natural F_0 contour of a specific syllable (i.e., with the smallest sum square error) is selected as its pitch target. An example of such resynthesis is shown in figure 11.7, where the short dashed lines are the learned targets. The F_0 contours generated with these learned targets (solid curves) seem to fit the natural F_0 contours (dotted curves) quite well. In phase 2, categorical targets are obtained by averaging over the parameters of all the syllables in the corpus that belong to the same categorical combination, for example, all the on-focus

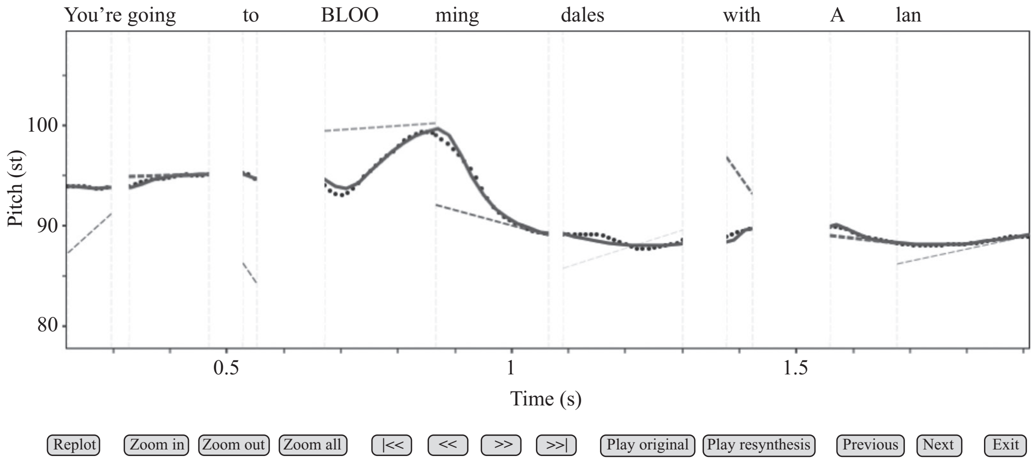


Figure 11.7
Original (dotted) versus resynthesized (solid) F_0 contours of the English utterance "You're going to Bloomingdales with Alan" by qTAtrainer. Pitch is F_0 given in semitones.

H tones that occur at the beginning of a sentence (Prom-on, Xu, and Thipakorn 2009). As found in Prom-on, Xu, and Thipakorn (2009) and Liu et al. (2013), good predictive results can be obtained for both English and Mandarin with this approach.

The categorization by averaging strategy employed in qTAtrainer, despite its reasonable performance, cannot satisfactorily estimate all qTA parameters. In particular, locally estimated parameters may not be globally optimal. For example, in some cases, the rate of target approximation (λ) may not be adequately estimated if there is severe target undershoot. In addition, the exhaustive search implemented in qTAtrainer is inefficient and probably ecologically unrealistic as a learning algorithm. These problems are addressed by PENTAtrainer, in which function-specific targets are learned from an entire corpus that has been functionally annotated (Prom-on and Xu 2012; Xu and Prom-on 2014). This is achieved with *simulated annealing*, an optimization algorithm that performs stochastic parameter sampling to avoid local minima in parameter estimation. Figure 11.8 shows an example of an annotated utterance (top) and natural F_0 and synthetic contours (bottom), where the latter is generated with categorical targets learned from an entire corpus.

In Xu and Prom-on (2014), good overall numerical results were achieved with PENTAtrainer for English (the same data set tested with qTAtrainer in Liu et al. 2013), Mandarin, and Thai. In Prom-on, Xu, and Thipakorn (2009), which applied categorization by averaging, the perceptual identification rates for tone in Mandarin and focus in both Mandarin and English were similar for synthetic and natural speech. Just as importantly, synthetic prosody (in terms of F_0 and duration) was heard to be as natural as natural prosody for English and only slightly worse for Mandarin.

The total number of function-specific parameters that need to be learned from the speech corpora to perform predictive synthesis is fairly small. In Xu and Prom-on (2014), seventy-eight parameters (twenty-six b, m, λ values each) were used for 960 English sentences (consisting of 8,640 syllables), eighty-four parameters for 1,280 Mandarin sentences (consisting of 10,240 syllables), and thirty parameters for 2,500 Thai disyllabic phrases. This suggests that a high level of abstraction can be achieved with

PENTA-based computational approaches. The abstraction level is comparable to other models, for example, five parameters per Standard Chinese tone in the Fujisaki model (Fujisaki 1983) and four parameters per intonational event in the Tilt model (Taylor 2000).

11.3.2 Modeling Encoding Schemes of English Prosody: An Illustration

The application of the computational tools described has allowed us to model some of the major prosodic encoding schemes in English and Mandarin. Figure 11.9 provides a summary illustration with modeling data on English from Xu and Prom-on (2014). Each graph shows the original F_0 of an American English utterance, pitch targets learned by PENTAtrainer, and synthetic F_0 contours generated with the learned targets. These sentences were spoken with either sentence-medial or sentence-final focus and as either statements or questions (Liu et al. 2013). As can be seen, the predicted F_0 contours closely simulate the interactive realization of multiple encoding schemes of lexical stress, focus, and modality in American English (as described in Xu and Prom-on 2014). These include, first, the robust post-focus pitch range shift, whose direction depends on modality: downward in a statement (a, b), but upward in a question (c, d). Second, it has also simulated the interactive determination of target slope by lexical stress, focus, and modality: negative slope for on-focus word-final stressed syllables in a statement, but positive in a question (*job* in b, d).

Note also that the match between the synthetic and original F_0 contours in figure 11.9 is not nearly as good as that in figure 11.7. This is partly because here, the synthesis is based on categorical parameters learned from all the utterances by a speaker in a corpus, as opposed to resynthesis in figure 11.7 (by qTAtrainer), but partly also because there is still room for further adjustments in our functional annotations. For example, because the relative position of unstressed syllables within an initial-stressed word is not annotated in this simulation, the pitch targets of the unstressed syllables are the same regardless of their positions in the word. As a result, the synthetic F_0 in *-crosoft* does not show the final upstep in figure 11.9d. Thus, even if the major characteristics of the encoding schemes have been identified, their detailed properties are still an object of continuous investigations.

11.4 Broader Significance

PENTA is a conceptual framework for characterizing speech prosody as a process of articulatorily encoding communicative functions. Thus, its central concern is how prosody operates as part of the speech communication process. At each stage of the development of PENTA, especially during its quantification, we have tried to ensure the operability of the model, that is, the ability to take input data in a sufficiently detailed form and generate outputs that are sufficiently close to surface acoustics, based on mechanisms that are biomechanically plausible. This operability is not just about how the model works internally, but also about how it is linked to external processes. As a result, PENTA is relevant not only for the direct characterization of prosody, but also for understanding many broader issues, including speech acquisition, language change, typology, and phonological representation. As will be shown in the following discussion, the implications for all these issues are most clearly seen through computational modeling.

(a)

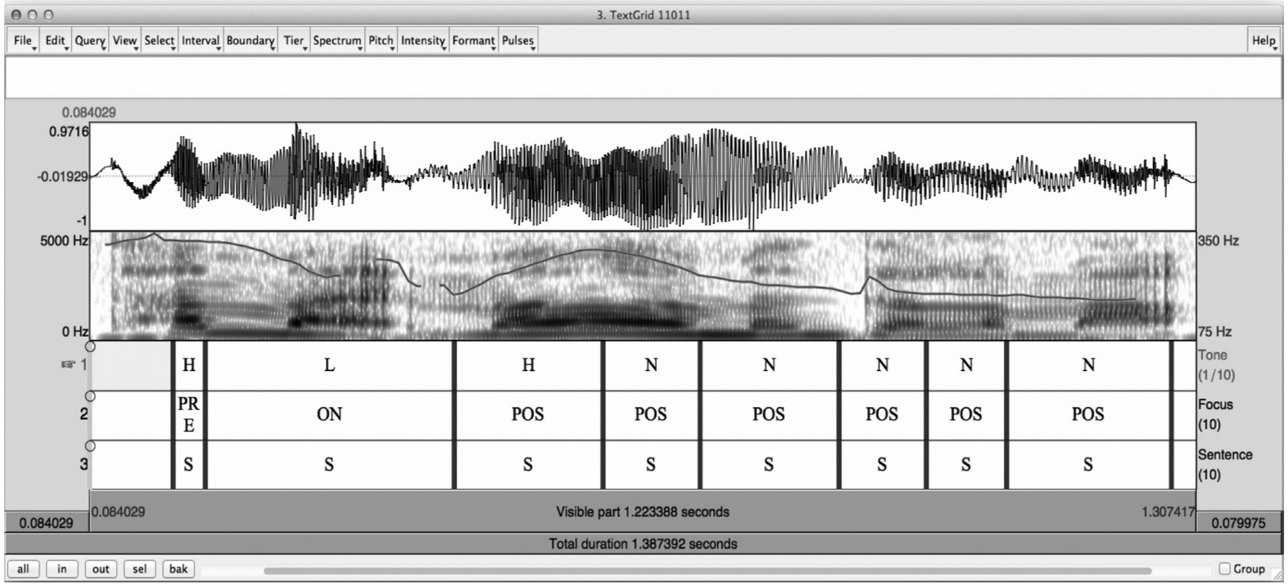
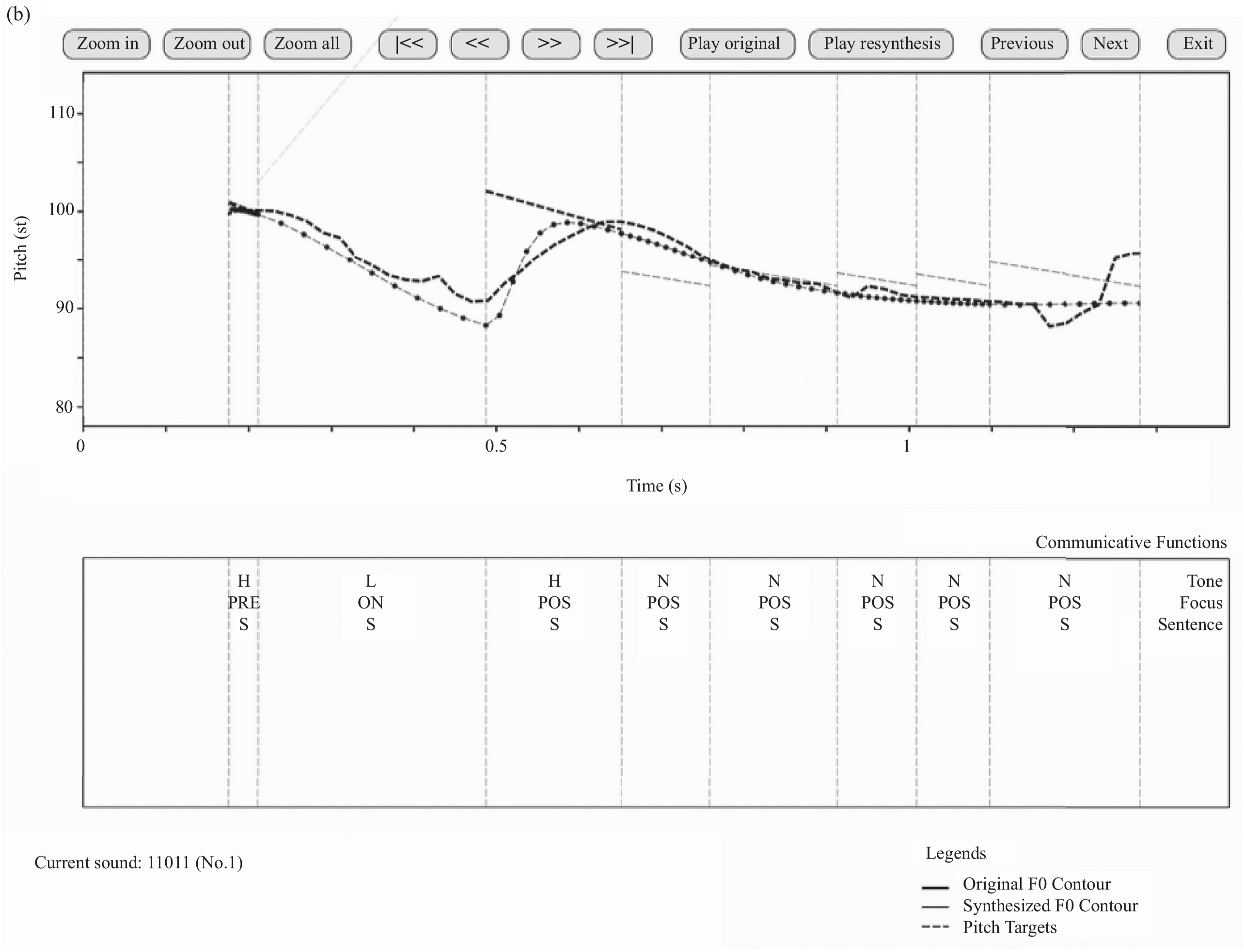

3. TextGrid 11011
File
Edit
Query
View
Select
Interval
Boundary
Tier
Spectrum
Pitch
Intensity
Formant
Pulses
Help
0.084029
0.9716
-0.01929
-1
5000 Hz
0 Hz
350 Hz
75 Hz
1
2
3
H
L
H
N
N
N
N
N
PR
E
ON
POS
POS
POS
POS
POS
POS
S
S
S
S
S
S
S
S
Tone
(1/10)
Focus
(10)
Sentence
(10)
0.084029
0.084029
Visible part 1.223388 seconds
1.307417
0.079975
Total duration 1.387392 seconds
all
in
out
sel
bak
Group

Figure 11.8
Screenshots of PENTAtrainer (www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer/) interfaces. The annotation interface (top) allows users to mark functional units and their temporal domains. In this example (Mandarin sentence "tā MĂI māma men de la ma?" [Did he BUY what mother has?], with focus on mai3), the annotated functions are lexical tone, focus, and sentence type (modality). All the boundaries are set to coincide with syllable boundaries. The temporal domain of a functional region covers syllables with identical labels. The output interface (bottom) displays learned pitch targets (dashed lines), synthetic (dotted), and natural (solid) F_0 contours. It also allows users to play the utterance with either synthetic or natural prosody (Prom-on and Xu 2012).


(b)
Zoom in
Zoom out
Zoom all
|<<
<<
>>
>>|
Play original
Play resynthesis
Previous
Next
Exit
Pitch (st)
110
100
90
80
0
0.5
1
Time (s)
Communicative Functions
H PRE S
L ON S
H POS S
N POS S
N POS S
N POS S
N POS S
N POS S
Tone
Focus
Sentence
Legends
Original F0 Contour
Synthesized F0 Contour
Pitch Targets

Current sound: 11011 (No.1)

Figure 11.8
(continued)

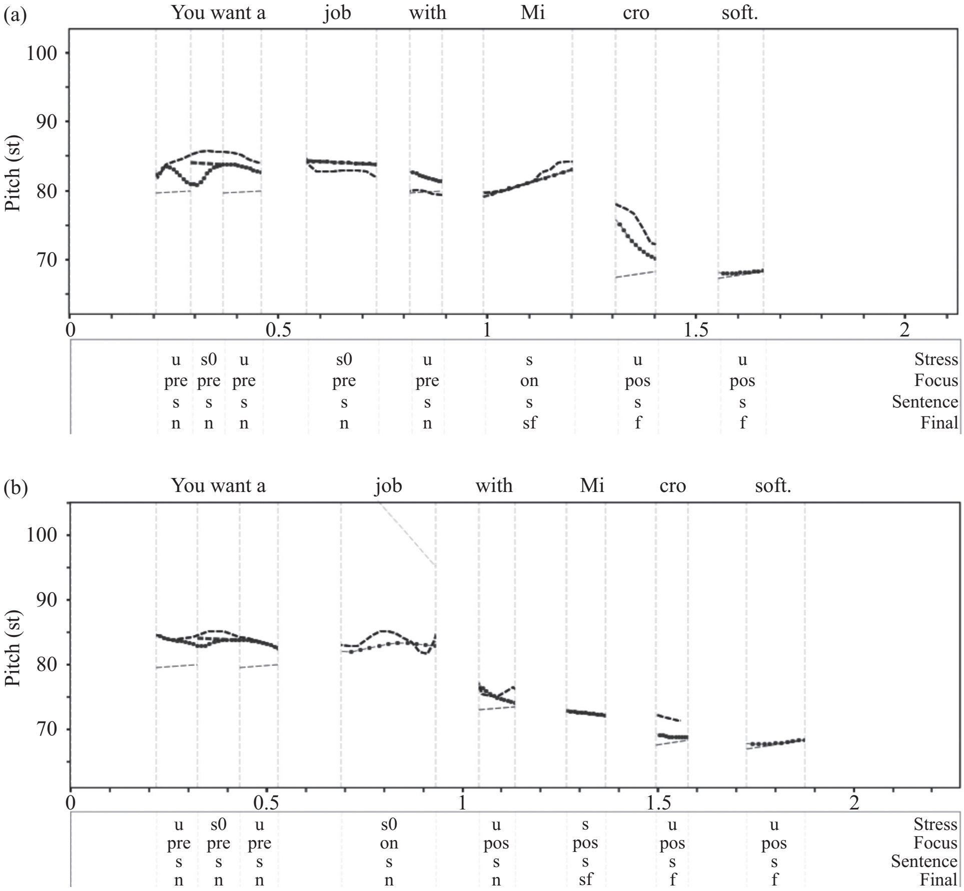


Figure 11.9
Original (dashed) and synthetic (dotted) F_0 contours of the sentence *You want a job with Microsoft*, spoken by a male American English speaker as either a statement (a, b) or a question (c, d), with focus on either *job* (b, d) or *Microsoft* (a, c). Also displayed are the pitch targets (straight dashed lines) learned by PENTAtrainer based on the functional annotations shown at the bottom of each graph. Stresses are categorized as unstressed (u), nonfinal stressed (S), and word-final stressed (s0). Syllable positions (labeled as final) are categorized as nonfinal (n), semifinal (sf), and sentence-final (f). All the graphs are screenshots of the demo window of the Synthesis tool (synthesize.praat) in the PENTAtrainer package. *Source*: Data courtesy of Xu and Prom-on (2014).

11.4.1 Role of Computational Modeling

The explanatory power of a scientific theory can be measured by the number of falsifiable predictions it can make and the level of specificity of these predictions. A theory of speech prosody, for example, can be evaluated in terms of the number of prosodic patterns it can predict and how closely the predicted surface forms match the natural ones in fine detail. A computational model, especially if it is able to generate continuous surface acoustic patterns, would offer an effective means of testing the predictive power of its corresponding prosodic theory. In contrast, a theory with no quantitative implementation is difficult to falsify, because it cannot generate predictions that are

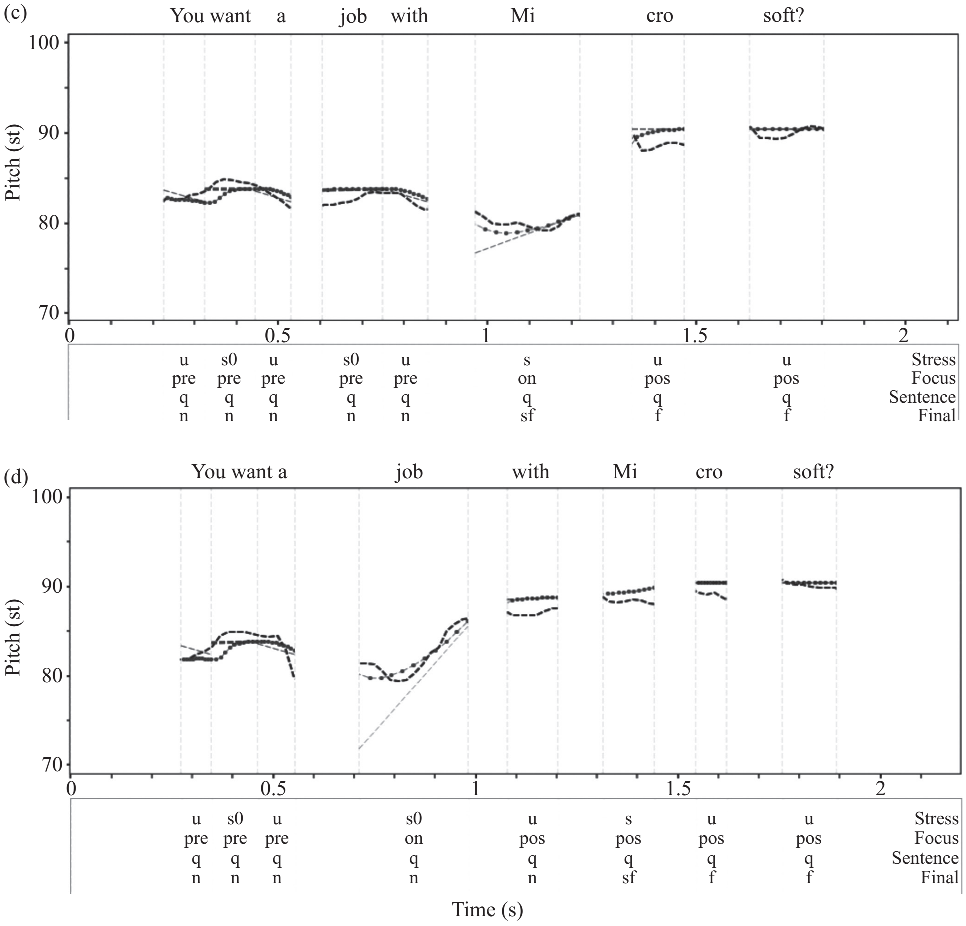


Figure 11.9
(continued)

detailed enough to be compared with the naturally occurring patterns or with other theories that do have quantitative implementations.

In practice, however, models can be constructed at different levels and generate predictions in different degrees of details. Theories such as autosegmental-metrical (AM) phonology and optimality theory are models that take certain forms of data (underlying segments or features) as input and generate predictions as output (narrow phonetic transcription) based on hypothetical mechanisms. None of these theories, however, generate outputs that are detailed enough for numerical acoustic comparisons with real speech. In other words, they lack adequate predictive power. For certain theories, for example, the AM theory of intonation (Ladd 2008; Pierrehumbert 1980; Pierrehumbert and Beckman 1988; Pierrehumbert and Hirschberg 1990), it is not even clear what the input is to the model-internal prosodic grammar. This, plus the fact that the output is largely symbolic, makes the theory hard to falsify.

Hence, computational modeling with fine prosodic details is not just a fancy and dispensable way of doing prosody. Modeling at a sufficient level of detail allows us to see the exact consequences of each assumption we make in a theory. Also, both speakers and listeners have to deal with fine phonetic details. So, any theory that claims to be relevant about the cognitive process of prosody eventually has to have an account of how prosody is articulated by adults and how it is acquired by children. Models that leave these details to theory-external processes cannot be considered cognitively operable, unless plausible interfaces with the external processes have been explicitly proposed.

11.4.2 Articulation as Part of Cognition

The preceding discussion also brings up an issue that so far has received little attention. It is generally assumed that anything cognitive happens only in the brain, and that motor movements, because they are external to the brain, are not part of cognition (Pierrehumbert 1990). What we have learned from computational modeling with qTAtrainer and PENTAtrainer suggests, however, that this may not be the case. Because the qTA model is a simulator of the articulatory process of pitch generation, extracting underlying pitch targets with both trainers mimics the use of articulators as part of a cognitive learning process. In PENTAtrainer, for example, during target extraction, the articulatory system represented by qTA is used to repeatedly generate F_0 contours with randomly selected pitch targets, and the contours are then compared to the natural ones. If the match is good enough, the target is accepted for the current stage of learning. If the match is bad, the hypothetical target is discarded and the random target selection continues. In this kind of learning, a biomechanical device (here simulated by qTA) is indispensable. This is because, without it, the brain would have to replicate all the peripheral processes, which would lead to the kind of paradox discussed in section 11.2.1.

More interestingly, what we have learned from modeling actually matches well with what we now know about vocal acquisition by songbirds and human children. Like human babies, young (male) songbirds have to learn their songs from adult males. But their learning takes a number of stages. In an early sensory learning stage, they need to hear the adult songs and store the songs as templates. Otherwise they would never learn to sing properly (Brainard and Doupe 2002; Petkov and Jarvis 2012). In the later sensorimotor learning stage, they no longer need to hear the adult songs, but need to hear their own song practices (Brainard and Doupe 2002). Deprived of such opportunity, they would again never sing properly. There is also evidence that human children follow a similar order of bootstrapped learning. Acquisition of normal speech is severely affected not only in children who lose hearing at a very young age, but also in those who are unable to vocalize during a later practice stage (Cowie and Douglas-Cowie 1992). And children who cannot generate or hear their own voice prior to puberty experience severe delays in their speech development (Doupe and Kuhl 1999; Kamen and Watson 1991; Kuhl and Meltzoff 1996; Locke and Pearson 1990).

Thus vocal learning in both songbirds and humans requires the use of the very articulatory system that is later used in mature song or speech production. Both the naturally occurring biological evidence and modeling evidence, therefore, point to the use of a biomechanical system as part of the cognitive process of vocal learning. This in turn demonstrates the usefulness of computational modeling in our theoretical understanding of speech, provided that biological plausibility is taken into full consideration during the development of the models. It also shows the relevance of simulating learning as a crucial task of computational modeling. After all, mature vocalization is but a behavior at the later stage of an incessant vocal learning process. Thus, demonstration of learnability should go a long way toward demonstrating full operability.

11.4.3 What about Phonology?

Another critical issue to be addressed in this chapter is whether PENTA sees a role for a phonological level of representation in prosodic theory. More specifically, does PENTA posit a set of abstract, symbolic primitives that can be combined to generate well-formed intonation contours linked to both meaning and a range of acceptable phonetic instantiations?

This issue goes back to the earlier discussion of the Saussurean view on linguistic symbols as unities of the signifiers and the signified. Because the symbols concerned in the earlier days of linguistics are structures like words and morphemes, it is natural to take it for granted that the symbols are relatively certain while the meanings are relatively vague. But the difficulty encountered in the study of prosody suggests that it is crucial to consider as early as possible the meanings conveyed by prosody, given the obscurity of the prosodic forms. This has led to our view that it is the signified, namely, the communicative functions, that should be the defining properties of prosodic units, while the signifiers, that is, the prosodic forms, should be empirically discovered for each of the communicative functions. This is the basis of the notion of encoding schemes in PENTA.

Although the encoding schemes are not the same as phonological entities such as pitch accents, phrase accents, and boundary tones (Pierrehumbert 1980), they do share some similarities. First, like phonological entities, encoding schemes are abstract, as they do not exactly resemble what is observable in surface prosody. Unlike phonological entities, however, encoding schemes are not symbolic, because they do not rely on symbolic values to define their identity.[3] Rather, their identity is defined by the respective communicative functions. Also, unlike pitch accents and boundary tones, encoding schemes are not the primitive units in PENTA. Rather, the prosodic primitives are the syllable-bound pitch targets. Their primitive status comes from the fact that they are obligatory in the production of the syllable, as discussed in section 11.2.1. But pitch targets are each associated to multiple functions, which makes them inappropriate as phonological units.

Most importantly, to PENTA, it is critical to identify the kind of representation that enables the transmission of communicative meanings from the speaker to the listener, that is, representations that are operable. An operable representation needs to be sufficiently abstract so as not to require too much memory resource. It also needs to be able to account for fully continuous surface forms, leaving as few details unexplained as possible. And it should allow adequate representation of individual and dialectal variations. Finally, it needs to be learnable with testable computational algorithms.

The solution found in the PENTA approach that can satisfy all of these requirements is parametric representation in the form of underlying articulatory targets. Table 11.1 shows a list of properties of such parametric representation as compared to symbolic representation. As can be seen, the parametric representation can satisfy all four requirements. First, it is abstract because it can represent an infinite number of contextual variants by translating underlying targets into surface acoustics based on model-simulated articulatory mechanisms. Thus, the representation itself is free of redundant surface details. It is also gradient, because the targets are numerically specified and so are not categorical themselves. This allows them to represent numerous

Table 11.1
Comparison of PENTA-based parametric and AM-style symbolic representations

Properties	Parametric	Symbolic
Abstract (able to represent multiple surface variants)	√	√
Gradient (allowing for individual and dialectal variations/representation itself is not categorical)	√	
Continuous (with built-in time-varying patterns)	√	
Data-driven (trainable, learnable)	√	?
Functionally defined	√	

dialectal and individual variations of the same linguistic categories. The targets are also continuous with built-in time-varying patterns. Given that the continuity is an intrinsic property of the target approximation model, the representation does not leave the filling of the detailed contours to other, unspecified mechanisms, as is the case with theories like AM. Furthermore, the specific values of this representation are data-driven, so they can be obtained from real speech data. More importantly, this way of parameter extraction allows the simulation of real-life learning of articulatory targets. This makes the representation computationally operable in terms of simulating biological reality with a reasonable level of plausibility.

Finally, even more directly relevant to the core concerns of phonology, the model-based parametric representation offers not only an operable way of linking meanings to articulatory targets, but also a biologically plausible mechanism for the emergence and evolution of rule-like phonological variations. This can be seen in two examples of our recent modeling study. The first one is about the highly perplexing phenomenon of tone sandhi (Chen 2000). As a well-known example, the Mandarin tone 3 assumes a T2-like surface form before another tone 3: T3 T2 / _ T3. From a functional perspective, such a "rule" makes little sense because it leads to homophony and so reduces categorical contrasts. But rules of this kind are commonplace in tone languages (Chen 2000), which is puzzling. So, there must be some strong biological constraint that makes the emergence of such communication-harming rules inevitable.

Our modeling simulation of tone and intonation with PENTAtrainer seems to offer a suggestion as to what this constraint might be. That is, it is probably because there is no other way of conveying communicative meanings than to load all the functions onto syllable-bound articulatory targets that are realized in succession. Because each target is co-specified by multiple functions, learners rarely encounter monofunctional targets. Thus, a tone is hardly ever learned as having a single-category target. Instead, for example, a specific version of T3 may be learned with a particular target when followed by another T3. If, for some unknown articulatory, perceptual, social, or historical reason, this version of T3 happened to sound more like T2 in surface form, a T2-like context-specific target could be learned for T3. But this context-specific T3 target does not need to be identical to the T2 target in the same context, because the functional combinations are not the same. As found in Xu and Prom-on (2014), the best modeling result was obtained when the sandhi T3 was allowed to learn its own target, rather than when it was forced to share the same target with T2. This result is in line with the empirical finding of subtle yet consistent differences between the original and sandhi-derived T2 in Mandarin (Peng 2000; Xu 1997) despite their full perceptual merger (Peng 2000; Wang and Li 1967)

Beside tonal context, many other frequently occurring functions can also be the conditioning factors, such as boundary marking, focus, sentence type, and so on. For example, as discussed in section 11.2.2, syllable-bound pitch targets are conditioned in American English by the interaction of lexical stress, focus, and sentence type (Liu et al. 2013). And, as shown in section 11.3.2, in PENTA-based computational modeling, these multifunctional targets can be easily learned from natural speech data (Xu and Prom-on 2014). The same principle is applicable to countless other cases. If these resemble biological reality, we would have identified a core mechanism behind many of the phonological rules.

11.4.4 Prosodic Typology

Through our empirical studies, the PENTA approach has been shown to be also relevant for prosodic typology. Specifically, typological phenomena, as we have found, are best described in terms of variations in the way specific communicative functions are encoded. The model itself, however, does not stipulate what prosodic form a language should take. Instead, it allows researchers to empirically discover for each function what its encoding is in a particular language. This function-oriented approach has led to our discovery of an interesting typological distribution of a prosodic pattern, namely, *focus*. Following a series of classic studies of focus realization in American English (Cooper, Eady, and Mueller 1985; Eady and Cooper 1986; Eady et al. 1986), we found that Mandarin Chinese shares a common prosodic feature with American English, which we termed *postfocus compression* (PFC). That is, the pitch range of postfocus words is lowered relative to the neutral-focus reference.

In later studies, we were surprised to find that Taiwanese (Southern Min) and Cantonese do not have this feature. From that, we hypothesized that this typological pattern is not accidental; rather, it is an indication that the origin of Mandarin and probably other northern Chinese dialects is different from that of southern Chinese languages such as Taiwanese and Cantonese. Furthermore, PFC has been found in a number of languages in language families that have been described as belonging to the putative Nostratic macro-family. We have therefore hypothesized that PFC originated from proto-Nostratic over thirteen thousand years ago (Bomhard 2008). One of the bases of this hypothesis is the finding that PFC is almost impossible to pass on from one language to another (Xu 2011a; Xu, Chen, and Wang 2012). It is also hard for it to emerge on its own, as none of the non-Nostratic languages examined for focus has shown evidence of PFC, regardless of whether they have lexical stress, tone, or any other prosodic features (Xu 2011a). PFC is also hard to acquire in a second language, as it seems to require the learner to speak the language more than their first language before PFC is consistently seen in their production (Chen 2015; Chen, Xu and Guion-Anderson 2014).

Another potential prosodic typology is about how focus interacts with sentence type. In American English, the pitch range of a postfocus word is raised well above the pitch median in questions. But this feature is missing in Mandarin (Liu and Xu 2005; Liu et al. 2013) and not reported in other languages. It is possible that both patterns are shared by many other languages, which is worth exploring in future studies.

Yet another typological divide is in terms of the interaction between focus and phrasing. In languages like English and Mandarin, focus can operate on the syllable level. When focus falls on a multisyllabic word, any syllable after the stressed syllable is treated as belonging to the postfocus domain (Xu and Xu 2005). In contrast, in languages like French, due to the constraint of phrase structure, or due to lack of lexical stress, auxiliaries do not show consistent pitch-accent–like patterns. In this case, the interaction of focus and phrasing is partially influenced by the lexical function, which differs across stress versus nonstress languages.

Finally, an even deeper typological divide has been suggested by Rialland (2009), who discovered that in a number of languages in a geographically restricted area in the Sudanic belt of Africa, prosodic means other than final pitch raising are used to indicate interrogation. The fact that the final pitch raising seems to be universal among languages outside Africa may suggest that this divide existed before *Homo sapiens* first left Africa over tens of thousands of years ago. All these patterns, if proven reliable by further studies, suggest that prosody is likely to be more stable than segmental, lexical, or even syntactic features during language change. If so, prosodic features could be used as indicators of language affinity that may have greater time depths than traditional indicators (Longobardi and Guardiano 2009; Nichols 1996).

11.4.5 Encoding Schemes as Prosophemes

In the lexical domain, the smallest meaning-bearing units are recognized as morphemes, which can be as small as a segment or as large as a multisyllable word or word root. In the prosodic domain, our work with PENTA has suggested that it is the encoding schemes that bear the closest resemblance to lexical morphemes (Liu et al. 2013). First, like lexical morphemes, each encoding scheme consists of multiple prosodic components. These components are meaningless by themselves, but they act jointly to mark both intra- and interfunctional contrasts. Second, similar to lexical morphemes, encoding schemes have allomorph-like variants whose occurrence is conditioned by factors such as location in sentence and interaction with other prosodic functions. Finally, similar to lexical morphemes, encoding schemes are language-specific, and their patterns likely have historical origins. Given that prosodic encoding schemes use prosody to carry postlexical meanings, we may call them *prosophemes*.

Our description of prosodic focus so far has made it a clear case of prosopheme. First, focus is realized not only with specific pitch patterns, but also with specific patterns of duration, intensity, and even voice quality (as reviewed in Xu, Chen, and Wang 2012). Second, PFC of pitch and intensity has been found to be language-specific, and languages with this feature have been hypothesized to be linked to a common protolanguage (Xu 2011a). The encoding scheme of focus in languages like Mandarin and English is therefore multicomponential, language-specific, and with likely historical etymologies, thus bearing all the major hallmarks of a lexical morpheme.

Another case is modality or sentence type, which encodes whether a sentence is a statement or yes/no question. In American English, for example, modality determines not only sentence-final F_0 (which is treated as boundary tone in AM theory), but also the underlying pitch targets of all stressed syllables throughout the sentence (Liu et al. 2013), as discussed in sections 11.2.2 and 11.3.2. Modality also interacts with focus to determine the pitch range of all postfocus words: well above the neutral-focus reference in a question, but well below the reference in a statement. Both of these features are missing in Mandarin. Question intonation in Mandarin does not involve postfocus raising of pitch range above the reference, and neither does it change the pitch targets of individual syllables (Liu and Xu 2005; Liu et al. 2013). So, again we see clear evidence of multicomponentiality, conditional variability, and language specificity in the encoding scheme of modality.

Therefore, the multicomponential coding of the prosodic functions demonstrates that it is the communicative functions, rather than the directly observable surface F_0

events, that bear the most resemblance to lexical morphemes. This prosopheme notion is an alternative to the tonal morpheme proposed by Pierrehumbert and Hirschberg (1990). As discussed in detail in Liu et al. (2013), many of the morpheme-like meanings proposed by Pierrehumbert and Hirschberg for the phonological intonational components are similar to those associated with prosodic functions such as focus and modality. Furthermore, some proposed phonetic implementation rules in AM theory (Pierrehumbert 1980; Pierrehumbert and Hirschberg 1990) are part of the morpheme-like characteristics of focus and modality. For example, the upstep rule in English, which is said to raise the portion of F_0 corresponding to a high boundary tone H% relative to the preceding H- phrase accent, is part of a continuous upshift of postfocus pitch range to mark a question. This extra raising is therefore morphophonological, that is, part of a prosopheme rather than being due to a phonetic implementation rule.

11.4.6 The Perceptual Perspective

So far we have had little discussion of how PENTA can account for the perception of prosody. A main reason is that we have not yet conducted extensive investigation of the perception of prosody, with the exception of studies on emotional prosody, which were mainly perception-based (Chuenwattanapranithi et al. 2008; Noble and Xu 2011; Xu, Kelly, and Smillie 2013; Xu et al. 2013). The few perceptual experiments we have performed on nonemotional prosody were mainly done with the purpose to verify the production patterns found in those studies (Taheri-Ardali, Rahmani, and Xu 2014; Xu, Chen, and Wang 2012; Xu, Xu, and Sun 2004) or to evaluate modeling performance (Prom-on, Xu, and Thipakorn 2009; Xu and Prom-on 2014). To test whether PENTA can also serve as a proper theory of prosody perception, especially in terms of being able to make precise predictions on how communicative functions are perceptually decoded from prosody, additional research is needed. Nonetheless, there is already evidence from modeling studies using self-organizing maps (Kohonen 1982) that the perception of tone and intonation is based on mechanisms that do not require full knowledge of production (Gauthier, Shi, and Xu 2007a, 2007b; Gauthier, Shi, and Xu 2009). This is consistent with the fact that perception learning precedes production learning in speech acquisition (Kuhl et al. 1992; Werker and Tees 2005). But more research, both behavioral and modeling, is needed to develop better predictive knowledge about prosody perception and how it is linked to the production of prosody.

11.5 Conclusion

PENTA was proposed based on the premise that prosody is a system of encoding communicative meanings with an articulatory system. From this basis, we have identified syllable-synchronized sequential target approximation as the core articulatory mechanism of prosodic encoding. In PENTA, therefore, syllable-bound pitch targets and their articulatory approximation are the phonetic primitives. On the meaning side, PENTA assumes that categories and dimensions of communicative meanings are defined by function rather than form. Thus, there are no theory-intrinsic units in PENTA that are equivalent to pitch accents, phrase accents, and boundary tones (Pierrehumbert 1980), or accent and phrase commands (Fujisaki 1983), or any units that are defined primarily by their phonetic properties. Yet the empirically established encoding schemes of communicative functions established under PENTA appear to bear close similarities to lexical morphemes, which gives rise to the term *prosopheme*. Empirically guided search for function-form unities has also led to our finding of prosody-based typological divisions such as those based on PFC, postfocus pitch upshift in questions, and differential interaction of focus and phrasing. The principle of parallel encoding in PENTA also makes it easy to incorporate more gradient and universal functions such as emotional codes based on the hypothetical bioinformational dimensions (Xu, Kelly, and Smillie 2013).

The quantization of PENTA has given us tools for computational modeling of prosody that can significantly increase the predictive power of the theory. Most interestingly, modeling has made it clear to us that the only way to acquire an encoding scheme is through the learning of syllable-bound multifunctional targets. Such a process, because it requires the learning of different targets for the same functional category conditioned by interactions with other functions, is a likely breeding ground for phonological rules such as tone sandhi that seem to make little sense from a purely functional perspective. Much more research is needed, nevertheless, to explore the full potential of computational modeling not only for speech prosody, but also for the segmental aspect of speech.

Notes

1. There is some overlap in content between this chapter and Xu et al. (2015).

2. This is facilitated by the software tool that we have developed, namely, ProsodyPro, a Praat-based script, available at www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/ (Xu 2013). Also, a similar tool for segmental analysis in the form of FormantPro (www.homepages.ucl.ac.uk/~uclyyix/FormantPro/). It facilitates systematic comparison of continuous formant trajectories, which is still rare in segmental studies.

3. Even if symbols sometimes are used in our approach, they are only for convenience of discussion. This also rules out PENTA as a transcription system because it does not use transcription as a means of analysis.

References

Bailly, G., and B. Holm. 2005. “SFC: A Trainable Prosodic Model.” *Speech Communication* 46:348–364.

Bernstein, N. A. 1967. *The Coordination and Regulation of Movements*. Oxford: Pergamon Press.

Bolinger, D. L. 1972. “Accent Is Predictable (If You’re a Mind Reader).” *Language* 48:633–644.

Bomhard, A. R. 2008. *Reconstructing Proto-Nostratic: Comparative Phonology, Morphology, and Vocabulary*. Leiden, the Netherlands: Brill.

Brainard, M. S., and A. J. Doupe. 2002. “What Songbirds Teach Us about Learning.” *Nature* 417 (6886): 351–358.

Chen, M. Y. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge: Cambridge University Press.

Chen, Y., and Y. Xu. 2006. “Production of Weak Elements in Speech: Evidence from F0 Patterns of Neutral Tone in Standard Chinese.” *Phonetica* 63:47–75.

Chen, Ying. 2015. “Post-focus Compression in English by Mandarin Learners.” In *Proceedings of the 18th International Congress of Phonetic Sciences*, edited by The Scottish Consortium for ICPhS 2015.

Chen, Y., Y. Xu, and S. Guion-Anderson. 2014. "Prosodic Realization of Focus in Bilingual Production of Southern Min and Mandarin." *Phonetica* 71:249–270.

Cheng, C., and Y. Xu. 2013. "Articulatory Limit and Extreme Segmental Reduction in Taiwan Mandarin." *Journal of the Acoustical Society of America* 134 (6): 4481–4495.

Chuenwattanapranithi, S., Y. Xu, B. Thipakorn, and S. Maneewongvatana. 2008. "Encoding Emotions in Speech with the Size Code: A Perceptual Investigation." *Phonetica* 65:210–230.

Connell, B., and D. R. Ladd. 1990. "Aspects of Pitch Realization in Yoruba." *Phonology* 7:1–29.

Cooper, W. E., S. J. Eady, and P. R. Mueller. 1985. "Acoustical Aspects of Contrastive Stress in Question-Answer Contexts." *Journal of the Acoustical Society of America* 77:2142–2156.

Cowie, R., and E. Douglas-Cowie. 1992. *Postlingually Acquired Deafness*. New York: Mouton de Gruyter.

de Saussure, F. 1916. "Nature of the Linguistics Sign." In *Cours de linguistique générale*, edited by C. Bally and A. Sechehaye, 66–70. New York: McGraw-Hill.

Doupe, A. J., and P. K. Kuhl. 1999. "Birdsong and Human Speech: Common Themes and Mechanisms." *Annual Review of Neuroscience* 22:567–631.

Eady, S. J., and W. E. Cooper. 1986. "Speech Intonation and Focus Location in Matched Statements and Questions." *Journal of the Acoustical Society of America* 80:402–416.

Eady, S. J., W. E. Cooper, G. V. Klouda, P. R. Mueller, and D. W. Lotts. 1986. "Acoustic Characteristics of Sentential Focus: Narrow vs. Broad and Single vs. Dual Focus Environments." *Language and Speech* 29:233–251.

Fujisaki, H. 1983. "Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing." In *The Production of Speech*, edited by P. F. MacNeilage, 39–55. New York: Springer-Verlag.

Gandour, J., S. Potisuk, and S. Dechongkit. 1994. "Tonal Coarticulation in Thai." *Journal of Phonetics* 22:477–492.

Gauthier, B., R. Shi, and Y. Xu. 2007a. "Learning Phonetic Categories by Tracking Movements." *Cognition* 103:80–106.

Gauthier, B., R. Shi, and Y. Xu. 2007b. "Simulating the Acquisition of Lexical Tones from Continuous Dynamic Input." *JASA Express Letters* 121:EL190–195.

Gauthier, B., R. Shi, and Y. Xu. 2009. "Learning Prosodic Focus from Continuous Speech Input: A Neural Network Exploration." *Language Learning and Development* 5:94–114.

Gay, T. J. 1968. "Effect of Speaking Rate on Diphthong Formant Movements." *Journal of the Acoustical Society of America* 44:1570–1573.

Hirst, D. J. 2005. "Form and Function in the Representation of Speech Prosody." *Speech Communication* 46:334–347.

Hombert, J.-M. 1977. "Consonant Types, Vowel Quality, and Tone." *Studies in African Linguistics* 8:73–190.

Hsu, C., and Y. Xu. 2014. "Can Adolescents with Autism Perceive Emotional Prosody?" In *Proceedings of Interspeech 2014*, edited by Haizhou Li, Helen Meng, Bin Ma, Eng Siong Chng, and Lei Xie, 1924–1928.

Hyman, L. M. 1993. “Register Tones and Tonal Geometry.” In *The Phonology of Tone*, edited by H. v. d. Hulst and K. Snider, 75–108. New York: Mouton de Gruyter.

Kamen, R. S., and B. C. Watson. 1991. “Effects of Long-Term Tracheostomy on Spectral Characteristics of Vowel Production.” *Journal of Speech and Hearing Disorders* 34:1057–1065.

Kelso, J. A. S. 1984. “Phase Transitions and Critical Behavior in Human Bimanual Coordination.” *American Journal of Physiology: Regulatory, Integrative and Comparative* 246:R1000–R1004.

Kelso, J. A. S., D. L. Southard, and D. Goodman. 1979. “On the Nature of Human Interlimb Coordination.” *Science* 203:1029–1031.

Kohonen, T. 1982. “Self-Organized Formation of Topologically Correct Feature Maps.” *Biological Cybernetics* 43:59–69.

Kuhl, P. K., and A. N. Meltzoff. 1996. “Infant Vocalizations in Response to Speech: Vocal Imitation and Developmental Change.” *Journal of the Acoustical Society of America* 100 (4): 2425–2438.

Kuhl, P. K., K. A. Williams, F. Lacerda, F., K. N. Stevens, and B. Lindblom. 1992. “Linguistic Experience Alters Phonetic Perception in Infants by Six Months of Age.” *Science* 255:606–608.

Ladd, D. R. 2008. *Intonational Phonology*. Cambridge: Cambridge University Press.

Laniran, Y. 1992. “Intonation in Tone Languages: The Phonetic Implementation of Tones in Yorùbá.” PhD diss., Cornell University.

Laniran, Y., and C. Gerfen. 1997. “High Raising, Downstep and Downdrift in Igbo.” (Presentation, *Seventy-First Annual Meeting of the Linguistic Society of America*).

Latash, M. L. 2012. *Fundamentals of Motor Control*. Orlando, FL: Academic Press.

Lee, A., S. Prom-on, and Y. Xu. 2017. “Pre-low Raising in Japanese Pitch Accent.” *Phonetica* 74 (4): 231–246.

Liu, F., and Y. Xu. 2005. “Parallel Encoding of Focus and Interrogative Meaning in Mandarin Intonation.” *Phonetica* 62:70–87.

Liu, F., Y. Xu, S. Prom-on, and A. C. L. Yu. 2013. “Morpheme-Like Prosodic Functions: Evidence from Acoustic Analysis and Computational Modeling.” *Journal of Speech Sciences* 3 (1): 85–140.

Liu, X., and Y. Xu. 2014. “Body Size Projection and Its Relation to Emotional Speech: Evidence from Mandarin Chinese.” In *Proceedings of Speech Prosody 2014*, edited by N. Campbell, D. Gibbon, and D. Hirst, 974–977.

Locke, J. L., and D. M. Pearson. 1990. “Linguistic Significance of Babbling: Evidence from a Tracheostomized Infant.” *Journal of Child Language* 17:1–16.

Longobardi, G., and C. Guardiano. 2009. “Evidence for Syntax as a Signal of Historical Relatedness.” *Lingua* 119 (11): 1679–1706.

Mechsner, F., D. Kerzel, G. Knoblich, and W. Prinz. 2001. “Perceptual Basis of Bimanual Coordination.” *Nature* 414:69–73.

Nichols, J. 1996. “The Comparative Method as Heuristic.” In *The Comparative Method Revised*, edited by M. Durie and M. Ross, 29–71. Oxford: Oxford University Press.

Noble, L., and Y. Xu. 2011. “Friendly Speech and Happy Speech: Are They the Same?” In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by Wai-Sum Lee and Eric Zee, 1502–1505.

Peng, S.-H. 2000. “Lexical versus ‘Phonological’ Representations of Mandarin Sandhi Tones.” In *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, edited by M. B. Broe and J. B. Pierrehumbert, 152–167. Cambridge: Cambridge University Press.

Petkov, C. I., and E. Jarvis. 2012. “Birds, Primates and Spoken Language Origins: Behavioral Phenotypes and Neurobiological Substrates.” *Frontiers in Evolutionary Neuroscience* 4: 12. https://doi.org/10.3389/fnevo.2012.00012.

Pierrehumbert, J. 1980. “The Phonology and Phonetics of English Intonation.” PhD diss., MIT.

Pierrehumbert, J. 1990. “Phonological and Phonetic Representation.” *Journal of Phonetics* 18:375–394.

Pierrehumbert, J., and M. Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pierrehumbert, J., and J. Hirschberg. 1990. “The Meaning of Intonational Contours in the Interpretation of Discourse.” In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Prom-on, S., F. Liu, and Y. Xu. 2012. “Post-Low Bouncing in Mandarin Chinese: Acoustic Analysis and Computational Modeling.” *Journal of the Acoustical Society of America* 132:421–432.

Prom-on, S., and Y. Xu. 2012. “PENTATrainer2: A Hypothesis-Driven Prosody Modeling Tool.” In *Proceedings of Exling 2012*, edited by A. Botinis, 93–100.

Prom-on, S., Y. Xu, and B. Thipakorn. 2009. “Modeling Tone and Intonation in Mandarin and English as a Process of Target Approximation.” *Journal of the Acoustical Society of America* 125:405–424.

Rialland, A. 2009. “African ‘Lax’ Question Prosody: Its Realisations and Its Geographical Distribution.” *Lingua* 119:928–949.

Silverman, K. 1986. “F0 Segmental Cues Depend On Intonation: The Case of the Rise after Voiced Stops.” *Phonetica* 43:76–91.

Stevens, K. N. 1998. *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Sundberg, J. 1979. “Maximum Speed of Pitch Changes in Singers and Untrained Subjects.” *Journal of Phonetics* 7:71–79.

Taheri-Ardali, M., Rahmani, H., and Y. Xu. 2014. “The Perception of Prosodic Focus in Persian.” In *Proceedings of Speech Prosody 2014*, edited by N. Campbell, D. Gibbon, and D. Hirst, 515–519.

Taylor, P. 2000. “Analysis and Synthesis of Intonation Using the Tilt Model.” *Journal of the Acoustical Society of America* 107:1697–1714.

Wagner, M. 2005. “Prosody and Recursion.” PhD diss., MIT.

Wang, C., Y. Xu, and J. Zhang. 2019. “Mandarin and English Use Different Temporal Means to Mark Major Prosodic Boundaries.” In *Proceedings of the 19th International Congress of Phonetic Sciences*.

Wang, W. S.-Y., and K.-P Li. 1967. "Tone 3 in Pekinese." *Journal of Speech and Hearing Research* 10:629–636.

Werker, J. F., and R. C. Tees. 2005. "Speech Perception as a Window for Understanding Plasticity and Commitment in Language Systems of the Brain." *Developmental Psychobiology* 46:233–251.

Xu, C. X., and Y. Xu. 2003. "Effects of Consonant Aspiration on Mandarin Tones." *Journal of the International Phonetic Association* 33:165–181.

Xu, Y. 1997. "Contextual Tonal Variations in Mandarin." *Journal of Phonetics* 25:61–83.

Xu, Y. 1998. "Consistency of Tone-Syllable Alignment across Different Syllable Structures and Speaking Rates." *Phonetica* 55:179–203.

Xu, Y. 1999. "Effects of Tone and Focus on the Formation and Alignment of F0 Contours." *Journal of Phonetics* 27:55–105.

Xu, Y. 2001. "Fundamental Frequency Peak Delay in Mandarin." *Phonetica* 58:26–52.

Xu, Y. 2005. "Speech Melody as Articulatorily Implemented Communicative Functions." *Speech Communication* 46:220–251.

Xu, Y. 2007. "Speech as Articulatory Encoding of Communicative Functions." In *Proceedings of the Sixteenth International Congress of Phonetic Sciences*, edited by J. Trouvain and W. J. Barry, 25–30.

Xu, Y. 2009. "Timing and Coordination in Tone and Intonation: An Articulatory-Functional Perspective." *Lingua* 119 (6): 906–927.

Xu, Y. 2011a. "Post-Focus Compression: Cross-Linguistic Distribution and Historical Origin." In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by Wai-Sum Lee and Eric Zee, 152–155.

Xu, Y. 2011b. "Speech Prosody: A Methodological Review." *Journal of Speech Sciences* 1:85–115.

Xu, Y. 2013. "ProsodyPro: A Tool for Large-Scale Systematic Prosody Analysis." In *Proceedings of Tools and Resources for the Analysis of Speech Prosody*, edited by B. Bigi and D. Hirst, 7–10.

Xu, Y. 2020. Syllable is a synchronization mechanism that makes human speech possible. *PsyArXiv*. https:// doi:10.31234/osf.io/9v4hr.

Xu, Y., S.-W. Chen, and B. Wang. 2012. "Prosodic Focus with and without Post-Focus Compression (PFC): A Typological Divide within the Same Language Family?" *Linguistic Review* 29:131–147.

Xu, Y., A. Kelly, and C. Smillie. 2013. "Emotional Expressions as Communicative Signals." In *Prosody and Iconicity*, edited by S. Hancil and D. Hirst, 33–60. Philadelphia: John Benjamins.

Xu, Y., A. Lee, S. Prom-on, and F. Liu. 2015. "Explaining the PENTA Model: A Reply to Arvaniti and Ladd (2009)." *Phonology* 32:505–535.

Xu, Y., A. Lee, W.-L. Wu, X. Liu, and P. Birkholz. 2013. "Human Vocal Attractiveness as Signaled by Body Size Projection." *PLoS ONE* 8 (4): e62397.

Xu, Y., and F. Liu. 2006. "Tonal Alignment, Syllable Structure and Coarticulation: Toward an Integrated Model." *Italian Journal of Linguistics* 18:125–159.

Xu, Y., and F. Liu. 2012. “Intrinsic Coherence of Prosodic and Segmental Aspects of Speech.” In *Understanding Prosody: The Role of Context, Function, and Communication*, edited by O. Niebuhr, 1–26. New York: Walter de Gruyter.

Xu, Y., and S. Prom-on. 2014. “Toward Invariant Functional Representations of Variable Surface Fundamental Frequency Contours: Synthesizing Speech Melody via Model-Based Stochastic Learning.” *Speech Communication* 57:181–208.

Xu, Y., and X. Sun. 2002. “Maximum Speed of Pitch Change and How It May Relate to Speech.” *Journal of the Acoustical Society of America* 111:1399–1413.

Xu, Y., and A. Wallace. 2004. “Multiple Effects of Consonant Manner of Articulation and Intonation Type on F0 in English.” *Journal of the Acoustical Society of America* 115 (part 2): 2397.

Xu, Y., and M. Wang. 2009. “Organizing Syllables into Groups: Evidence from F0 and Duration Patterns in Mandarin.” *Journal of Phonetics* 37:502–520.

Xu, Y., and Q. E. Wang. 2001. “Pitch Targets and Their Realization: Evidence from Mandarin Chinese.” *Speech Communication* 33:319–337.

Xu, Y., and C. X. Xu. 2005. “Phonetic Realization of Focus in English Declarative Intonation.” *Journal of Phonetics* 33:159–197.

Xu, Y., C. X. Xu, and X. Sun. 2004. “On the Temporal Domain of Focus.” In *Proceedings of International Conference on Speech Prosody 2004*, edited by K. Hirose, 81–84.

Yuan, J., M. Liberman, and C. Cieri. 2006. “Towards an Integrated Understanding of Speaking Rate in Conversation.” In *Proceedings of Interspeech 2006*, edited by R. M. Stern, 541–544.

Zemlin, W. R. 1988. *Speech and Hearing Science: Anatomy and Physiology*. Englewood Cliffs, NJ: Prentice Hall.

Commentary on Chapter 11: Comparing the PENTA Model to Autosegmental-Metrical Phonology

Janet B. Pierrehumbert

Introduction

In this chapter, Xu, Prom-on, and Liu provide an overview of the recent work on the PENTA (parallel encoding and target approximation) model of prosody. This is a third-generation model of prosody and intonation. In characterizing the model this way, I am taking classic works based on auditory transcriptions, such as Bolinger (1958), Trager and Smith (1951), and Crystal (1969), as first-generation models, and the autosegmental-metrical (AM) models launched in the 1970s and 1980s as second-generation models.

Second-generation models benefited enormously from the rise of computer workstations with specialized software for speech processing, which enabled researchers to examine thousands of F0 contours and to create experimental stimuli in which melodic characteristics of speech were varied in a controlled manner. However, the developers of AM models did not yet have the inference and optimization methods that have played such a central role in the development of PENTA. Generative AM algorithms, such as those laid out in Anderson, Pierrehumbert, and Liberman (1984) and Beckman and Pierrehumbert (1986), were very seat-of-the pants efforts, compared to the multiparameter trajectories that are achieved with PENTA. So, we can ask: Has PENTA superseded the AM approach? To what extent has it built on insights of the earlier approach? What aspects of the AM research program remain even now as topics for future research?

A central issue in addressing these questions is the comparison between prosody and word phonology. AM theory, a development within generative linguistics, had as one of its goals a unified formalism for describing segmental, rhythmic, and melodic aspects of speech at both the lexical and the phrasal levels. AM theory adopted from phonemic theory the claim that each language has a relatively small inventory of phonological units, that these are combined with each other in accordance with language-specific constraints (which define the phonological grammar of the language), and that the mapping between form and meaning has a great deal of arbitrariness. *Go* is a verb of motion in English, but a board game in Japanese. *Puck* is a disk used in ice hockey, but with the change of a single feature, the word is obscene.

This arbitrariness, in combination with the relative simplicity of phonology, means that the phonological level defines a bottleneck between the phonetic realizations of words and their meanings. This bottleneck helps the language learner to acquire a large vocabulary by allowing articulatory and perceptual patterns exhibited in one word to be reused in other words. It promotes fast, accurate speech perception because words that are highly confusable from a phonetic point of view are unlikely to be similar in meaning, and thus unlikely to be lexical competitors in any given context.

These benefits accrue precisely because the flow of information from the extremely rich, high-dimensional world of meanings and communicative functions to the phonetics is severely restricted. (See Beckman and Pierrehumbert 2000 for further discussion.) Insofar as intonational patterns resemble word forms, then, they should be describable in terms of combinations of phonological units that could convey different meanings in some other combinations. Pierrehumbert and Hirschberg (1990) is an effort to provide a semantics for intonational patterns that has such a structure. Xu, Prom-on, and Liu's essay (this volume) represents a fundamental departure. Starting from a negative assessment of the extent of convergence in the field about the units of analysis, it proposes direct mappings from prosodic or intonational meanings (*communicative functions*) to phonetic parameters. This mapping does not have the same modules as those assumed in AM theory.

AM differed from much earlier work in phonology, such as classical phonemic theory or Chomsky and Halle (1968), in positing units that are phonetically realized at many different timescales. This development was a response to observations about the relationship of sound structure to meaning. The job of phonology is to describe *systems of contrast*—differences in form that speakers can use to communicate differences in meaning. It became evident that some contrasts are available in fewer positions than others. For example, it is very common for languages to support a full set of consonantal contrasts in syllable onset position, but only a restricted set in syllable coda position. English and Russian both have a larger set of vocalic contrasts in stressed syllables than in unstressed ones. In analyzing African tone languages, Leben (1971) and Goldsmith (1976) found that, while only vowels could carry distinctive tones, the number of distinctive tones in a word did not correspond to the number of vowels. Instead, the number of melodies for an entire word in languages such as Mende or Hausa is quite restricted, and in long words, a tone is spread out over two or more syllables. This observation naturally leads to the suggestion that the melody is underlyingly a characteristic of the word (and not of the phoneme or syllable), with general principles of tonal mapping intervening between the underlying form and the phonetic outcome. Intonation patterns involve an even larger timescale, as many intonational distinctions can be made only once per phrase. While syllables typically have durations in the range of one hundred to two hundred milliseconds, depending on their stress and position in the phrase (Arai and Greenberg 1997; Dankovicova 1999), intonation phrases are typically longer. The shortest intonation phrases are monosyllables that are among the longest syllables (being stressed and phrase-final), while phrases with two to four content words plus any associated clitics are more common, often taking a second or more to produce. A central goal of AM theory is to articulate what contrasts are available at what temporal scales.

In pursuit of this goal, AM theorists proposed units and principles for combining them. Some of the most widely accepted units are the segment, the mora, the syllable, the metrical foot, and the intonational phrase. Essentially all AM grammars had at least one unit in between the metrical foot and the intonational phrase, such as the prosodic word or the accentual phrase, but the name and nature of this unit varied with the research group and the target language. For the theory of prosody and intonation, it is obvious that a timescale larger than the segment is needed, because distinctive tonal contrasts are sparser than segmental contrasts in every language that has been studied.

AM phonology also asked whether the inventory of units is universal, whether tonal contrasts are assigned at the same temporal scale in all languages, and how the units used to describe melody and rhythm relate to those used in the lexical phonology.

Xu, Prom-on, and Liu's essay refers to three different temporal scales (the utterance, the word, and the syllable), although the paper does not take a clear stand on what sizes of units are available. The syllable is held to be a privileged universal unit for the planning and execution of speech output, while some communicative functions are selected at a much larger timescale, such as the utterance.

To summarize, then, their essay has fewer levels of representation for prosody and intonation than AM theory does. It relates communicative functions directly to phonetic outcomes. AM theory asserts one set of relations between meanings and the phonological representation, and a second and qualitatively different set of relations between the phonological representation and the phonetic outcome. Furthermore, their essay has a less elaborated theory of the units of sound structure, assigning a role to the syllable that is more privileged than its role in AM theory.

Semantics versus Phonology

The flat representational apparatus of PENTA entails that prosodic meanings have qualitatively different behavior from word meanings. For word meanings, there is widespread agreement that an intermediate level of representation, in between semantics and phonetics, is needed. This agreement is based on the dissociations observed between phonetic similarity and semantic similarity. By defining an intermediate level of representation, two distinct systems of relations are made available in the model. One associates meanings with phonological forms, while the other associates phonological forms with parametric outcomes. The phonological level reduces the dimensionality of the system to be learned. It acts as a bottleneck, blocking arbitrary mappings of semantic features to phonetic parameters. For example, features of meaning such as spiciness, permanency, or biodegradability have no direct claims on F0 or voice quality. To express these meanings, a speaker must select a suitable word, and then access the phonological representation of the word, whose limited properties in turn control the phonetic realization. In PENTA, it is claimed that communicative functions are mapped directly to quantitative parameters. The framework has two levels of representation instead of three. The PENTA encoding system, described as being similar to morphemes, has the job of capturing the relationship of these two levels. By using only a single set of relations, PENTA is claiming that the prosodic system lacks the properties that have historically motivated an intermediate level of representation for the segmental makeup of words. To assess this claim, we can ask to what extent intonational meanings do, or don't, resemble the meanings of words.

Probably the easiest example of natural language semantics is imageable nouns: a word form for an imageable noun is associated with the set of objects or events that provide examples of the relevant category. Despite the undeniable existence of some ideophonic words, in general the similarities among the forms and referents of nouns are poorly correlated, a property of linguistic systems referred to as *duality of patterning*. *Spat* has no particular semantic relationship to nouns made by minimal changes in its form (e.g., *spam*, *scat*). In sound, *pine* might be confusable with *tine*, but in meaning, a pine is more similar to a cedar or a spruce than to a tine. Other types of words have more elusive meanings that are very discourse-dependent. Words like *indeed* and *even* are interpreted with regard to the space of possible events or states of affairs that the speaker assumes to be in the listener's mind. However, words with such slippery semantics still provide plenty of examples of dissociations between form and meaning. For example, the three intensifiers *really*, *super*, *awfully* are each more similar in sound to other semantically unrelated words (e.g., *lily*, *tuber*, *Aussie*) than they are to each other. That is, they share a communicative function, asserting that something has some characteristic to a high degree, while differing greatly in form.

In languages with lexical tones, these participate in similar dissociations. They can be identified by finding minimal pairs that differ only in their tone, but refer to unrelated concepts. The well-known *ma* set of Chinese even provides a minimal quadruplet. This behavior is what leads Beckman and Pierrehumbert (1986) to treat Japanese, traditionally characterized as a "pitch accent language," as a tone language with a sparse assignment of tones. Two otherwise identical words with unrelated meanings may differ only in the presence or absence of a high (H)-low (L) contour in the lexical representation. In languages such as English, where the location of lexical stress can vary, similar observations can be made about stress. *Coral*, *vermillion*, and *tangerine* are very similar colors, but they have the main stress in different positions: initial, medial, and final position (with *tangerine* actually having two stresses). The trochaic pattern of *coral* is the single most common stress pattern of English, meaning that words with this stress pattern have no shared communicative functions.

In short, lexical tone and lexical stress act like phonological elements in that they are poorly correlated with semantic dimensions of meaning, while being highly predictive of the phonetic realization of a word. Lexical tone plays an important part in determining the F0 contour and voice quality, while lexical stress has complex ramifications for duration, force of articulation, and timing. In PENTA, lexical tone and lexical stress are described as *lexical communicative functions* because they distinguish words from each other. By this standard, all of segmental phonology should also be included in the "lexical communicative functions." For example, the phonological feature [+voice] distinguishes *pin* from *bin* and *coat* from *goat*. However, Xu, Prom-on, and Liu explicitly deny that the communicative function level corresponds to the phonological level, instead drawing a parallel between the PENTA encoding schemes and morphemes.

The focal communicative function may at first appear to be a more promising candidate for direct mapping from function to phonetic outcome. I will discuss focus in detail because it is both a key example of a communicative function in PENTA and a central topic in semantic and pragmatic theory. A starting point for the semantic treatment of focus is the fact that it can change the truth conditions for sentences, just as substituting one word for another can change the truth conditions for a sentence. An example from Rooth (1992) illustrates this fact:

(1) I only said that Carl likes HERRING.
(2) I only said that CARL likes herring.

Sentence (1) is false if I also said that Carl likes some relevant food other than herring, such as kippers. Sentence (2) is false if I also said that some other relevant person, such as Sam, likes herring. These differences are related to the differing presuppositions of the sentences. Sentence (1) presupposes I said that Carl likes something. Sentence (2) presupposes I said that somebody likes herring. The instantiations of "something" and "somebody," respectively, are produced with narrow focus. The many other words whose contribution to sentential semantics depends on the focus placement include *even*, *not*, *which*, and *always*.

PENTA would not have serious trouble with these particular examples. In the model, focus is treated as a unified communicative function, by which novel information is made phonetically prominent and other information is backgrounded. In these examples, the novel information and the phonetic prominence coincide. But is this the case in general?

It turns out that the locations of novel information and of phonetic prominence do not always coincide. In AM theory, focus provides one of the key arguments for a modular theory distinguishing the semantic/pragmatic, phonological, and phonetic levels of representation. The focus feature +F is a semantic feature, associated with new information that instantiates variables in the presuppositions of a sentence. The accentedness within the phrase is a phonological feature, associated with phonetic prominence. The association +F with the locations of accents is rather complex.

One complication arises from the existence of examples in which given material is accented. In a study of the Boston Radio News corpus, Shattuck-Hufnagel, Ostendorf, and Ross (1994) show that the locations of prenuclear accents are determined mainly by phonological constraints and not by information status, and Gussenhoven (1999) also argues that prenuclear accents are not semantically informative in English. The issue is not confined to prenuclear accents, however. Schwarzschild (1999) discusses cases such as (3). Here capitalization is used to indicate the nuclear accents, and one of these, on the second instance of MOON, falls on given material; indeed its entire phrase is given:

(3) The rising of the TIDES % depends on the MOON being full, % and the MOON being full % depends on the position of the SUN %.

A production study by German, Pierrehumbert, and Kaufmann (2006), following up on observations by Ladd (1980), shows that speakers often place the nuclear accent on a given noun in preference to a new sentence-final preposition. These same cases also show that new information can fail to be accented. Finally, the theory also needs to cover cases in which focused information is old, but still displays the characteristic behavior of focus in defining the scope of a semantic operator. This situation is illustrated in (4).

(4) Speaker A: Everybody already knew that Mary only eats VEGETABLES.
Speaker B: If even PAUL knew that Mary only eats vegetables %, why didn't he suggest a different restaurant?

Here, the nuclear accent in speaker B's first clause falls on *Paul*, and *vegetables* is deaccented, even though it is the focus for *only*. As shown by Beaver et al. (2007), weak phonetic reflexes of the +F feature support the idea that focus is present, but they are very slight compared to those for a first-occurrence focus with nuclear stress.

Further complications arise from the fact that nuclear accent on a word is interpreted in some cases, but not others, as a focus marking for an entire phrase containing that word. This phenomenon, the contrast between the "narrow focus" and "broad focus" interpretations of an utterance, displays a strong dependence on syntactic factors, as in the study by Birch and Clifton (2002). Their study explores how nuclear accents map differentially to broad focus, depending on whether the accented word is a syntactic head, an argument, a modifier, or an adjunct. Vallduvi and Engdahl (1996) documents the ways in which the interaction of focus with syntactic constraints can be language-specific.

Such dissociations between the semantic feature +F and the location of the nuclear accent in the phrase provide a classic argument for an intermediate level of representation. In view of the role of syntactic factors, I should say "at least one" intermediate level of representation, but here I will concentrate on the phonology. Words that are emphasized even though they are given, such as the second occurrence of MOON in (3), are produced with heightened prominence. In English, words that follow them in the same intonational phrase (here, "being full") are subordinated. Though the assignment and expression of focus remains an active research area, all competitive approaches to the issue follow Schwarzschild (1999) in using a constraint satisfaction architecture in which the accent placement in any individual phrase results from the interaction of semantic focus with additional factors stated at other levels of linguistic representation.

Because PENTA does not impose any restrictions on the number of communicative functions, nor on the formal power or complexity of mappings between communicative functions and phonetic outcomes, it is in principle capable of describing this rich variety of phenomena. For example, in response to German, Pierrehumbert, and Kaufmann (2006), the PENTA authors could split apart the communicative functions of "focus on a noun" and "focus on a preposition." In the case of a final preposition, the encoding function could push the phonetic prominence back onto the preceding noun. This approach would miss significant generalizations. It would make it appear coincidental that the outcome is pretty much the same as having focus on the noun itself. It would miss the fact that the phonological and syntactic entities that are alluded to in a more standard treatment are independently needed to capture many other linguistic patterns. A more conservative response would be for the PENTA authors to allow that the "communicative function" of focus in PENTA is actually a phonological feature—the location of the nuclear accent in the phrase. This response would pave the way for PENTA to be integrated with modern linguistic theories of focus and accent placement that cover a wider range of phenomena.

A well-established communicative function of intonation that Xu, Prom-on, and Liu do not discuss is the scalar implicature. Constructions that involve implicit scales provide one of the most studied areas of semantics and pragmatics. Lexical choices and intonational choices both provide key examples. So I will begin with some examples of lexical choices. The meanings of commonplace adjectives such as *tall*, *good*, *hot* are rooted in the discourse structure, because each makes a claim about the position of the modified noun with regard to a contextually evoked scale (Kennedy 2007). In the examples in (5), the standards for *tall*, *good*, and *hot* are relativized to the scale for the relevant comparison set. Sentence (5b) is not contradictory, because the relevant dimensions along which goodness is assessed differ for jack-o-lanterns and pies.

(5) a. Max is tall for an American, but short for a basketball player.
b. Howdon Biggy pumpkins are good for jack-o-lanterns, but poor for pies.
c. 100 Celsius is hot for a sauna, but not for an oven.

Now if *warm* means having a relatively high temperature, what does a sentence like (6) mean?

(6) It's warm.

The speaker relies on mutual knowledge of a contextually relevant scale (temperatures of saunas? temperatures of ovens?) and then she also implies that the temperature is lower than hot on that scale: If it really was *hot*, it would have been more cooperative and informative to say so. Thus, if the sauna is heating up, (6) would be appropriate if the temperature is 70 degrees Celsius but not yet 100, but if the oven is heating up, the sentence might be used if the temperature is 120 but not yet 180.

In English intonational phonology, the interpretation of the rise-fall-rise contour, transcribed as L*+H L-H% in the tones and breaks indices (ToBI) system (Silverman et al. 1992), also depends on the contextually relevant scale of comparison. A naturalistic study (Ward and Hirschberg 1985), a theoretical analysis (Pierrehumbert and Hirschberg 1990), and a perception experiment (Hirschberg and Ward 1992) all converge on the conclusion that speakers use this contour to implicate uncertainty about the scale or a scalar value. The speaker may be genuinely uncertain or may implicate uncertainty to express irony or disbelief. The following example, which I used as part of a classroom exercise developed with Mary Beckman, illustrates this communicative function:

(7) Speaker A: Great America lets you ride the Vortex only if you're over 5′3″.
Speaker B: MELANIE L*+H rides the Vortex, and she's 5′2″.

Here, speaker A evokes a scale of acceptability as a Vortex passenger, based on height. Speaker B expresses their reservations about the scale by challenging the entailment that Melanie falls too low. He uses the L*+H accent to underline this reservation. In our classroom exercise, we have found that beginning linguistics students reliably associate such rejoinders with a phrase expressing disbelief ("Really?") as opposed to a word expressing agreement ("Indeed!"). "Indeed!" is perceived as consistent with an H* accent. This example provides a good example of a pragmatic meaning that can be conveyed with a word choice, an intonational choice, or both. Other well-known examples of such meanings include turn-taking (does the speaker mean to hold the floor, or yield it?) and topic shifts. These can also be signaled with words (*Now*, . . . *On the other hand* . . .), with manipulations of the pitch range and accent choice, or with combinations of these (Hirschberg and Litman 1993).

Such observations mean that when the speaker wishes to convey any of these meanings, she makes a selection among the means available. Some combinations of meanings and forms of expression are mutually exclusive. For example, the speaker cannot end a sentence with two different words simultaneously. Example (8) provides an illustration of mutual exclusivity in the intonational domain:

(8) Are you coming or NOT H* L L%.

This is a yes/no question, and according to Xu, Prom-on, and Liu's essay, its communicative function should be conveyed with a rising F0 pattern in English. However, the sentence is also an exhaustive disjunction. In (8), which has the most typical intonation for this construction (Pruitt and Roelofsen 2013), the falling terminal pattern indicates that the list is complete. Its use supersedes the option of using a rising pattern to indicate that the truth value of the proposition under discussion is not known, and input from the listener is expected (Pierrehumbert and Hirschberg 1990). Because of the phonological bottleneck provided by the small number of tonal elements available at the end of the phrase, the speaker cannot simultaneously signal a complete list and an open proposition.

Because the encoding functions in PENTA are so unrestricted, and indeed are designed to handle cases in which gestures associated with different functions seem to be overlaid, PENTA says nothing about how some choices may be mutually exclusive. AM theory, in contrast, specifies a particular set of bottlenecks, including, for each language, the inventory of contrastive pitch accents or lexical tones, the constraints on metrical structure, and restrictions on the pileup of contrastive tonal events at phrase boundaries. These have the effect of limiting how many distinct dimensions of meaning can be conveyed by the prosody and intonation of any given expression.

I have sketched some ways in which lexical tone and intonation resemble segmental features in the lexicon, and the relation of focus to nuclear accent points to a distinction between the semantic feature +F and the phonological feature of nuclear accent. I've also suggested that there is significant overlap between prosodic and intonational meanings, on the one hand, and the meanings that words have, on the other hand. It is not clear, however, that prosodic and intonational meaning are really as arbitrary as what we see for segmental phonology. This question is difficult to address, particularly in the light of recent work on lexical iconicity and the emergence of conventions in sign languages (Meir et al. 2013; Dingemanse et al. 2015; Dingemanse et al. 2016). However, Xu, Prom-on, and Liu and AM theorists are in agreement that prosodic and intonational systems are language-specific and therefore must be learned. For example, a key reference in Xu, Prom-on, and Liu's essay, Rialland (2009), documents the association between low tone and questioning in a group of African languages, the opposite of what is found in many other language families. Although it is frequently assumed that stressed syllables universally are marked with high tones, this is not the case for Indic languages such as Bengali (Hayes and Lahiri 1991). Nor does the high tone have a constant meaning even within English. Depending on where a high tone (or F0 peak) occurs, it may be associated with a contrastive emphasis (the L + H* accent), a yes/no question (the terminal H%), a scalar implicature (the L* + H discussed), or a change of topic (see Hirschberg and Litman 1993).

Phonology and Phonetic Realization

Acknowledging that tones are specified more sparsely than segmental features, PENTA takes the syllable to be the domain of tonal specification. A substantial body of research in phonology and phonetics indeed treats the syllable as a privileged unit for language acquisition, motor planning, and speech perception (MacNeilage 1998). AM theory has not, in general, viewed the syllable as privileged to this extent. In AM theory, the syllable is one of a hierarchy of different prosodic units controlling contrasts at different timescales, and these interact to determine the F0 contour for any given time span. This hierarchy is motivated by assessment of contrastiveness, on the one hand, and by different timescales for phonetic realization, on the other.

The component of AM theory that deals with contrastiveness is covered by the rubric *prosodic licensing*. The general idea is that phonological representations include a structural skeleton, and different pieces of the content are associated with different parts of the skeleton. In much the same way, a house is a structure with an entrance and an optional stairwell; a doorknocker must go with an entrance, and a banister goes with a stairwell. Each structural position thus provides an opportunity to make some set of choices: A doorknocker or not? A high vowel or a low vowel? An offglide or not? Admittedly sometimes the choice is a Hobson's choice, as in the case of Model T Fords (structures that were always painted black) and the Tokyo Japanese accentual phrase (which always begins with an L tone). But in any case, if the rate at which some feature is selected is relatively rapid, then the feature is licensed by some smaller unit. If it is specified less often, then it is licensed by a bigger unit. Units that do not in themselves license a feature may still carry information about it that can be exploited in speech perception. Xu, Prom-on, and Liu defend the choice of the syllable as the tone-bearing unit in PENTA in part on the grounds that unstressed syllables in postfocal position have different contours than do stressed syllables in English, indicating the presence of F0 targets where AM theory does not posit any pitch accents. This is an interesting phenomenon. Previously, Grice, Ladd, and Arvaniti (2000) reported that the phrase accent in the ToBI framework can dock on a postnuclear stressed syllable instead of at the phrase boundary, leading to differences between stressed and unstressed syllables in postnuclear position. Beaver et al. (2007) observe a marginal effect of secondary focus in postnuclear position on the F0 range and minimum F0 of the focused syllable. Such observations indicate that phonetic realization systems such as that of Anderson, Pierrehumbert, and Liberman (1984) are overly simplistic. They do not, however, demonstrate that the postfocal stressed syllables support an independent choice of contour type. Insofar as the postnuclear stressed syllables display traits that are licensed at the phrasal level, an AM theorist might compare the situation to seeing a reflection of the banister in the hall mirror. The banister is a property of the stairway. If there is a hall mirror, it can indirectly reveal what the banister looks like, but it is not possible for the decorator to make distinct selections for the banister and the reflection of the banister in the mirror.

The concept of licensing is applicable both in the lexical phonology and in the phrasal phonology. In many languages, the set of consonantal contrasts available in the coda position is smaller than in onset position. For example, in the middle of the word in Japanese, the coda consonant (if any) must either be a nasal sharing the same place of articulation as the following onset or be the first part of a geminate. This means that the coda position licenses only the feature [+/– nasal] (Ito, Mester, and Padgett 1995).

In English, the full range of vowels is found only in stressed syllables in a foot. Unstressed syllables do not license all the vowel features. But the metrical foot is also the minimal domain for a pitch accent. This means that if a phrase is just one word with two metrical feet, such as *tangerine*, it can have both a nuclear and a prenuclear accent. But a sentence such as *It's a plane*, when produced with all the function words cliticized (and therefore unstressed), has only one accent. In the African tone languages I already discussed, it is not possible to make a fresh choice of tone with every syllable; a limited number of melodies are assigned to the whole word. In Japanese, the word is also the domain for the pitch accent; a word either has an HL melody or it doesn't. Such licensing constraints mean that Mende, Hausa, and Japanese contrast with languages such as Thai, in which every syllable may have its own tone. At both the lexical and the phrasal levels, the edges of domains may have a special status for licensing. While English has rather restricted consonant clusters within monomorphemic words, extra coronal obstruents are allowed at the end of the word. This gives us word-final consonantal pile-ups, as in *fifths* and *folds*. Similarly, the end of an intonation phrase provides an opportunity for an extra tonal event. In the middle of the phrase, it is impossible to have so much F0 movement on a single syllable. As a result, the presence of a complex contour can provide a cue to listeners that a phrase boundary exists, as shown in Streeter (1978).

The licensing system and the phonetic requirements are not always well-aligned. Misalignments can occur because the licensed elements are too dense, in relation to the speed of implementation that is possible from an articulatory point of view. They can also occur because the elements are too sparse, failing to specify some phonetic properties that are essential for the speech to be produced. Both of these discrepancies are addressed in AM theory and in PENTA. The approaches have some similarities and differences, and both approaches leave some problems unresolved.

If a feature is licensed at a very high density, the realization of one value may overlap that of a preceding or following value. A much-studied case of this situation is nasal coarticulation, in which lowering of the velum for a nasal consonant begins during a preceding non-nasal segment. The exact timing and extent of this effect differ from language to language, providing an important case of language-specific detailed phonetic learning (Beddor and Krakow 1999). In the domain of tone and intonation, Bruce (1977) already drew attention to the articulatory undershoot found in Swedish when a phrasal tone is crowded with a lexical accent. For English, Anderson, Pierrehumbert, and Liberman (1984) uses linear smoothing to approximate some consequences of tonal crowding. PENTA builds on the insights of these AM papers in its treatment of tonal crowding, but goes much further with its approach to characterizing the detailed F0 dynamics.

However, an important unsolved problem in AM theory for handling tonal crowding also remains in PENTA. As shown in Silverman and Pierrehumbert (1990), tonal crowding may be ameliorated by temporal displacement of tones instead of by undershoot (an in-depth discussion of the issue may be found in the same volume: Bruce 1990). The extent and interaction of these two options are not fully elucidated in AM theory nor in PENTA and may await a more explicit motor theory of tonal production. A particular difficulty for PENTA in handling tonal displacement as a solution to crowding is the claim (see section 11.2.1) that the inertia of the articulators only comes into play at one level; there is no representation of inertia in the brain. This claim appears to preclude the use of feedforward motor control strategies that are known to be acquired with intense practice and that appear necessary to model language-particular segmental coarticulation patterns (see Rosenbaum 2010 for a review of closed loop, open-loop, and feedforward motor control models.)

Xu, Prom-on, and Liu also admit that PENTA does not yet handle cases in which more than one distinctive tonal target is assigned to a single syllable. This occurs in AM theory when a pitch accent and one or more boundary tones are all assigned to a monosyllabic utterance, such as *Anne L*+H L H%*. By virtue of its licensing theory, AM theory defines equivalence classes among the F0 contours of utterances with very different numbers of syllables. *Anne L*+H L H%*, *Melanie L*+H L H%*, and *Melanie L*+H did it L H%* all count as having instances of the same contour type, with the scalar implicature discussed herein. PENTA does not make this equivalence because it does not yet cover the shortest utterances with complex contours.

If the licensed features are misaligned with the phonetic requirements because they are too sparse, AM theory provides two different responses. One is categorical copying or spreading at the phonological level. The other is phonetic interpolation, where the time in between fully specified locations is taken up by making transitions with more or less alacrity.

Turkish vowel harmony provides an example where the first approach is justified; the harmonizing features that are lacking on the vowels in the underlying representation of a suffix are spread from the stem. In the nonconcatenative Arabic morphological system, both vowels and consonants can be copied to fill in all the structural positions supplied by the morphological pattern. In languages with reduplication, it is common for an affix to be merely a structural template, which is filled by copying material from the stem. The original AM treatments of African tone languages similarly asserted that tones were spread or copied to supply F0 targets for otherwise toneless vowels, as vocalic articulations require a specification of the source. The second approach is taken in Keating's (1988) treatment of the tongue position for the phoneme /h/. The tongue position differs greatly between /ihi/ and /aha/, because /h/ lacks intrinsic oral specifications such as [high] or [front]. In Anderson, Pierrehumbert, and Liberman (1984), Silverman et al. (1992), and Pierrehumbert and Beckman (1988), the option of underspecifying syllables for tones is heavily exploited. For both English and Japanese, according to these models, the alacrity of transitions between contrastive tonal specifications depends systematically on how far separated these specifications are. If people have plenty of time to make the transition, they adjust the articulators more slowly. The status of some of these analyses seems clear-cut, but not all. The unrestricted character of phonetic realization rules in AM theory sometimes makes it difficult to distinguish between phonological and phonetic analyses of the same pattern. Notably, in the Hobson's choice cases I alluded to, nothing really prevents the suggestion that the entire F0 contour is supplied by the phonetic realization component, without any phonological representation of the melodic features at all.

PENTA avoids the vexed distinction between phonological copying or spreading and phonetic realization by denying the existence of a separate phonological level. In cases where a communicative function is sparsely specified (e.g., focus is assigned to only one word per phrase), but has wide-ranging effects (both the focused word, and neighboring words, are affected), the encoding of the communicative function calculates a contour for each syllable in the affected sequence. The syllable has a special status for planning the output, but does not have the same status in the set of communicative functions, which are specified at the syllable, word, and phrasal levels, depending on the function.

The encoding manipulates several gradient parameters. It shares with Fujisaki and Hirose (1984) and Garding (1987) the claim that local gestures (such as lexical tones) interact with larger-scale planning, which has its own dynamics. However, unlike these works, PENTA does not superimpose local contours on a phrasal contour. Instead, phrasal choices (e.g., focus and modality) induce modulations of the pitch range in a manner that is reminiscent of Liberman and Pierrehumbert (1984) and Pierrehumbert and Beckman (1988). Like Kochanski and Shih (2003), the strength or weakness of different gestures, representing the strength of the requirement to achieve the specified target, is also gradiently manipulated. Very broadly speaking, the PENTA encoding builds on the pitch range control mechanisms proposed in AM, while also building on insights from other work that support more detailed modeling of the F0 dynamics. Thanks to its method for optimizing the parameters for detailed representations of F0 contours, it can achieve much more reliable matches to natural F0 contours than AM did and accordingly a better quality in speech synthesis. Overall, PENTA's strong reliance on gradient parameters would cause AM theorists to view it as a new and particularly sophisticated example of a phonetic realization model. This view is reinforced by PENTA's agnostic view about the system of underlying contrasts.

Generalization

Generative phonology, like generative linguistics, has the goal of predicting the form of examples that were not used in developing the model. Capturing the capabilities of native speakers of a language, the grammar of the language should be able to predict and analyze the forms of previously unobserved phrases. Predictions about language typology should also follow from varying the parameters in the overall formal framework.

For AM theory, the primary test of within-language generalization was intonation synthesis for novel phrases and sentences. A number of intonation synthesis algorithms using the framework were demonstrated and used commercially. Contrary to the claim of Xu, Prom-on, and Liu, several AM models do generate F0 contours that are detailed enough for comparisons with real speech. For example, Pierrehumbert and Beckman (1988, chapter 6) provides exact equations for synthesizing F0 contours of Japanese, and such contours can be compared with those in recordings. Pitrelli and Eide (2003) describes a system in which ToBI labels are associated with communicative functions of questioning and contrastive emphasis. Decision trees were then statistically trained to map these features together with text features into F0 and duration patterns. This system achieved good results in a perceptual evaluation using novel sentences.

Validation of PENTA is presented in Xu and Prom-on (2014) and cited in the present work. This 2014 paper presents details of the corpora used to train PENTA and cross-validate the parameter estimates. The data set is a highly constrained set of read text designed to elicit contrasts in focus location and in questions versus statements through the use of prompt sentences. The parameters are trained on all but one of the speakers in the data set and then used to predict the F0 contours for the held-out speaker. This procedure tests generalization across speakers at the parametric level, a dimension of generalization that is little explored in AM theory.

Validating models on held-out test data is the standard in other areas of speech and language engineering, and the application of this method in prosody and intonation is welcome. However, it is also important to recognize the limitations of this training and validation procedure. First, the prompt sentences represent a minimal discourse context, in comparison to the longer contexts that have led to the discoveries of some communicative functions of intonation that were mentioned herein. The corpus does not include the full range of intonational patterns that are found in expressive, colloquial speech. Second, optimization algorithms that attempt an overall match to a data set are very vulnerable to frequency effects, with detailed fits for the more frequent cases often achieved at the expense of coverage of less frequent cases. This can lead to problems with robustness and adaptation in new situations where the frequencies of the different cases may be quite different. This vulnerability was addressed by constructing a corpus with balanced representation of all the functions covered by the model, undercutting the suggestion that the learning method used in PENTA resembles that used by the cognitive system. By means that are not yet fully understood, humans can learn from highly unbalanced training sets and also adapt to different situations. Finally, because all speakers read the same materials, the study does not evaluate the model's ability to generalize to previously unseen sentences.

AM theory generated claims about language typology. The approach predicted that languages could differ in many independent dimensions. Questions for the phonological grammar of any given language included:

Underlying differences

- How many tonal elements does the language have? What are the basic tonal units that recombine in complex contours?
- Are tones part of the representations of words in the lexicon?
- Which phonological units license tones, whether syllables, feet, or larger units?
- Which tones are assigned at the phrasal (postlexical) level?
- Which semantic/pragmatic meanings are associated with which tonal sequences?

Realizational differences

- How is the realization of each tone affected by its metrical and tonal context?
- How are the tones licensed by each phonological unit aligned with the segments licensed by the same unit?
- What happens when word-level and phrase-level tones pile up on short material, such as monosyllabic words?

Such questions separate factors that were lumped together in the traditional distinction among tone languages, pitch-accent languages, and intonation languages. The result was extensive research on the typology of prosody and intonation. Studies established cross-linguistic differences on essentially all dimensions specified by the theory, culminating in works such as Jun (2005).

PENTA has also been used for cross-linguistic comparisons, with the discussion in the target paper focused on two dimensions of variation. One is the presence versus absence of postfocus compression, which is found in English and Mandarin but not in Taiwanese. The other is the use of low rather than high pitch to mark interrogation, as documented in Rialland (2009). These comparisons lead Xu, Prom-on, and Liu to suggest that prosody and intonation are exceptionally stable, with patterns possibly being preserved for time periods over ten thousand years. This conclusion is at odds, however, with other studies that have found systematic differences among dialects that diverged much more recently. Pierrehumbert and Beckman (1988) note differences in the implementation of low tones between Tokyo and Osaka Japanese, Bruce (2004) finds seven distinct intonational dialects of Swedish, and intonational variation among dialects of English has also been extensively investigated (Fletcher, Grabe, and Warren 2005; Grabe and Post 2002). Such differences can be discovered because it is possible to identify them using AM theory, with its tool kit of different factors at different levels of representation. The striking persistence of some prosodic and intonational patterns, together with the recent origins of others, provides another point of resemblance to the rest of phonology, in which long-standing patterns and rapid changes are also found (Coleman 2016).

Conclusion

In comparing PENTA and AM theory, a central issue is how much prosody and intonation resemble words in the lexicon. I have argued that there are considerable resemblances. The meanings of intonational contours are similar to those of many words.

Word choices and intonational choices affect the truth values of sentences. They also reflect pragmatic factors, such as the universe of comparisons in a discourse context. In some cases, speakers have a choice of communicating their meaning with a word choice, an intonation pattern, or both.

For words, dissociations between form and meaning motivate an intermediate level of representation, namely, the phonology. PENTA denies that this level of representation exists for prosody and intonation. This decision creates difficulties for the interpretability of the model. The data set on which the system was trained conflates the semantic feature of focus with the phonological feature of nuclear accent. These do not necessarily coincide. It is not clear how PENTA would handle examples of focused material that is not accented, or accented material that is not focused. PENTA treats stress and lexical tone as communicative functions, on a par with semantic focus and interrogation. However, these have no general semantic or pragmatic meanings; their function is differentiating words, which in classical linguistic theory is the hallmark of phonological elements. Thus, PENTA conflates the phonological description (at the lexical level) with the syntactic, semantic, and pragmatic description (at the level of the phrase or utterance). This means that the theory has not tackled many issues that were previously framed within AM theory and linguistic semantic theory. Overall, the communicative functions of PENTA resemble components of an underlying phonological representation more than they resemble semantic and pragmatic functions. These observations lead to the conjecture that PENTA is a novel and highly optimized theory of phonetic realization (e.g., of the mapping between the phonology and the phonetic outcome).

As a phonetic realization model, PENTA adopts some of the advances that AM theory made in the 1980s. It integrates information at different timescales, and it generates F0 contours for all syllables, including those with and without stress, or with and without lexical tone. The AM phonetic realization algorithms reflected both conceptual and technical limitations. Conceptually, they were influenced by a strong presumption that the underlying linguistic system is minimized, using descriptive elements that are as simple as possible and as orthogonal as possible. Since that time, it has been clear that the human cognitive system can learn very detailed patterns and often represents them with a great deal of redundancy. In this connection, the amount of quantitative detail in PENTA is perfectly plausible. In particular, the idea of tonal targets that have both a position in the pitch range and a dynamics is not problematic. The AM realization algorithms were also limited technically by the inability to optimize the values of the parameters that they did have. PENTA carries out an up-to-date and highly effective optimization. It also meets current standards by validating on held-out test materials. This means that it represents a substantial advance in our understanding of phonetic realization.

References

Anderson, M. J., J. B. Pierrehumbert, and M. Y. Liberman. 1984. “Synthesis by Rule of English Intonation Patterns.” In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 77–80. Piscataway, NJ: IEEE. https://doi.org/DOI 10.1109/ICASSP.1984.1172427.

Arai, T., and S. Greenberg. 1997. “The Temporal Properties of Spoken Japanese Are Similar to Those of English.” *Proceedings of Eurospeech* 2:1011–1014.

Beaver, D., B. Z. Clark, E. Flemming, T. Florian Jaeger, and Maria Wolters. 2007. “When Semantics Meets Phonetics: Acoustic Occurrences of Second-Occurrence Focus.” *Language* 83 (2): 245–276.

Beckman, M. E., and J. B. Pierrehumbert. 1986. “Intonational Structure in Japanese and English.” *Phonology Yearbook* 3:13–70.

Beckman, M. E., and J. B. Pierrehumbert. 2000. “Positions, Probabilities, and Levels of Categorization.” In Barlow, M. (Ed.) *Proceedings of the Eighth Australian International Conference on Speech Science and Technology*, Australian Speech Science and Technology Association, Canberra. 1–18.

Beddor, Patrice Speeter, and Rena Arens Krakow. 1999. “Perception of Coarticulatory Nasalization by Speakers of English and Thai: Evidence for Partial Compensation.” *Journal of the Acoustical Society of America* 106:2868–2887.

Birch, S., and C. Clifton. 2002. “Effects of Varying Focus and Accenting of Adjuncts on the Comprehension of Sentences.” *Journal of Memory and Language* 47:571–588.

Bolinger, Dwight. 1958. "A Theory of Pitch Accent in English." *Word* 14 (2–3): 109–149.

Bruce, G. 1977. *Swedish Word Accents in Sentence Perspective*. Lund, Sweden: Gleerup.

Bruce, G. 1990. "Alignment and Composition of Tonal Accents: Comments on Silverman and Pierrehumbert's Paper." In *Papers in Laboratory Phonology I*, edited by J. Kingston and M. E. Beckman, 107–114. Cambridge: Cambridge University Press.

Bruce, G. 2004. "An Intonational Typology of Swedish." In Isabelle (Eds). *Speech Prosody 2004*, Nara, Japan, March 23–26, 2004, edited by Bernard Bel and Isabelle Marlien,175–178. https://www.isca-speech.org/archive/sp2004/sp04_175.html.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper.

Coleman, J. 2016. "Acoustic-Phonetic Modelling of Historical and Prehistoric Sound Change." Paper presented at LabPhon15, Cornell University. July 13–16, 2016.

Crystal, David. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Dankovicova, J. 1999. "Articulation Rate Variation within the Intonation Phrase in Czech and English." In Proceedings of *ICPhS99*, University of California, Berkeley. 269–272.

Dingemanse, M., D. Blasi, G. Lupyan, M. H. Christiansen, and P. Monaghan. 2015. "Arbitrariness, Iconity, and Systematicity in Language." *Trends in Cognitive Sciences* 19 (10): 603–615.

Dingemanse, M., W. Schuerman, E. Reinisch, S. Tufvesson, and H. Mitterer. 2016. "What Sound Symbolism Can and Cannot Do: Testing the Iconicity of Ideophones from Five Languages." *Language* 92:e117–e133.

Fletcher, J., E. Grabe, and P. Warren. 2005. "Intonational Variation in Four Dialects of English: The High Rising Tune." In *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun, 390–407. Oxford: Oxford University Press.

Fujisaki, H., and K. Hirose. 1984. "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese." *Journal of the Acoustical Society of Japan* 5:233–242.

Garding, E. 1987. "Speech Act and Tonal Theory in Standard Chinese: Constancy and Variation." *Phonetica* 44 (1): 13–29.

German, J. S., J. B. Pierrehumbert, and S Kaufmann. 2006. "Evidence for Phonological Constraints on Nuclear Accent Placement." *Language* 82 (1): 151–168.

Goldsmith, John A. 1976. "An Overview of Autosegmental Phonology." *Linguistic Analysis* 2:23–68.

Grabe, E., and B. Post. 2002. "Intonational Variation in the British Isles." In *Speech Prosody 2002*, Aix-en-Provence, France, April 11–13, 2002, 275–278. https://www.isca-speech.org/archive/sp2002/sp02_275.html.

Grice, M., D. R. Ladd, and A. Arvaniti. 2000. "On the Place of Phrase Accents in Intonational Phonology." *Phonology* 17:143–185.

Gussenhoven, C. 1999. "On the Limits of Focus Projection in English." In *Focus: Linguistic, Cognitive and Computational Perspectives*, edited by P. Bosch and R. van der Sandt, 43–55. Cambridge: Cambridge University Press.

Hayes, B., and A. Lahiri. 1991. "Bengali Intonational Phonology." *Natural Language and Linguistic Theory* 9 (1): 47–96.

Hirschberg, Julia, and Diane Litman. 1993. “Empirical Studies on the Disambiguation of Cue Phrases.” *Computational Linguistics* 19 (3): 501–530.

Hirschberg, Julia, and Gregory Ward. 1992. “The Influence of Pitch Range, Duration, Amplitude and Spectral Features on the Interpretation of the Rise-Fall-Rise Intonation Contour in English.” *Journal of Phonetics* 20:241–251.

Ito, J., A. Mester, and J. Padgett. 1995. “Licensing and Underspecification in Optimality Theory.” *Linguistic Inquiry* 26 (4): 571–613.

Jun, S.-A. 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Keating, P. A. 1988. “Underspecification in Phonetics.” *Phonology* 5:275–292.

Kennedy, C. 2007. “Vagueness and Grammar: The Semantics of Relative and Absolute Gradable Adjectives.” *Linguistics and Philosophy* 30:1–45.

Kochanski, G., and C. Shih. 2003. “Quantitative Measurement of Prosodic Strength in Mandarin.” *Speech Communication* 41 (4): 625–645.

Ladd, D. R. 1980. *The Structure of Intonational Meaning: Evidence from English*. Bloomington: Indiana University Press.

Leben, W. 1971. “Suprasegmental and Segmental Representation of Tone.” In *Papers from the Second Conference on African Linguistics, UCLA, March 26–27. Studies in African Linguistics, Supplement 2*.

Liberman, M. Y., and J. B. Pierrehumbert. 1984. “Intonational Invariance under Changes in Pitch Range and Length.” *Language Sound Structure: Studies in Phonology Presented to Morris Halle*, edited by Mark Aronoff and Richard T. Oehrle, 157–233. Cambridge, MA: MIT Press.

MacNeilage, P. F. 1998. “The Frame/Content Theory of Evolution of Speech Production.” *Behavioral and Brain Sciences* 21 (4): 499–511.

Meir, I., C. Padden, M. Aronoff, and W. Sandler. 2013. “Competing Iconicities in the Structure of Languages.” *Cognitive Linguistics* 24 (2): 309–343. https://doi.org/10.1515/cog-2013–0010.

Pierrehumbert, Janet B., and Mary E. Beckman. 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pierrehumbert, J. B., and J. Hirschberg. 1990. “The Meaning of Intonational Contours in the Interpretation of Discourse.” In *Intentions in Communication*, edited by P. Cohen, K. Morgan, and M. Pollack, 271–311. Cambridge, MA: MIT Press.

Pitrelli, J. F., and E. M. Eide. 2003. “Expressive Speech Synthesis Using American English ToBI: Questions and Contrastive Emphasis.” In *IEEE Workshop on Automatic Speech Recognition and Understanding*. Piscataway, NJ: IEEE. https://doi.org/10.1109/ASRU.2003.1318524.

Pruitt, K., and F. Roelofsen. 2013. “The Interpretation of Prosody in Disjunctive Questions.” *Linguistic Inquiry* 44 (4): 632–650.

Rialland, A. 2009. “African ‘Lax’ Question Prosody: Its Realisations and Its Geographical Distribution.” *Lingua* 119:928–949.

Rooth, M. 1992. “A Theory of Focus Interpretation.” *Natural Language Semantics* 1:75–116.

Rosenbaum, D. A. 2010. *Human Motor Control*. Orlando, FL: Academic Press.

Schwarzschild, R. 1999. "AvoidF and Other Constraints on the Placement of Accent." *Natural Language Semantics* 7:141–177.

Shattuck-Hufnagel, S., M. Ostendorf, and K. Ross. 1994. "Stress Shift and Early Pitch Accent Placement in Lexical Items in American English." *Journal of Phonetics* 22:357–388.

Silverman, Kim, Mary E. Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. "ToBI: A Standard for Labeling English Prosody." In *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff, Alberta, Oct 12–16, 867–870. University of Alberta.

Silverman, Kim, and Janet Pierrehumbert. 1990. "The Timing of Prenuclear High Accents in English." In *Papers in Laboratory Phonology I*, edited by J. Kingston and M. E. Beckman, 107–114. Cambridge: Cambridge University Press.

Streeter, L. A. 1978. "Acoustic Determinants of Phrase Boundary Perception." *Journal of the Acoustical Society of America* 64(6): 1582–1592.

Trager, G. L., and H. L. Smith. 1951. *An Outline of English Structure*. Norman, OK: Battenberg Press.

Vallduvi, E., and E. Engdahl. 1996. "The Linguistic Realization of Information Packaging." *Linguistics* 34:459–519.

Ward, G., and J. Hirschberg. 1985. "Implicating Uncertainty: The Pragmatics of Fall-Rise." *Language* 61:747–776.

Xu, Y., and S. Prom-on. 2014. "Towards Invariant Functional Representations of Variable Frequency Contours: Synthesizing Speech Melody via Model-Based Stochastic Learning." *Speech Communication* 57:181–208.

Author Response to the Commentary: Multiple Layers of Meanings Can Be Linked to Surface Prosody without Direct Mapping

Yi Xu, Santitham Prom-on, and Fang Liu

PENTA Is Not a Direct Mapping Model

We are delighted to see Pierrehumbert's characterization of parallel encoding and target approximation (PENTA) as a third-generation model of prosody and intonation. Indeed, much of the refinement PENTA may potentially bring to our understanding of prosody has benefited from knowledge gained from empirical research since the earlier models. One of the key insights from empirical findings is that surface prosodic forms, such as F0 peaks, valleys, elbows, whole contours, and so on, cannot be mapped to underlying units, be it tone, stress, pitch accents, or prominence. This insight is instrumental in the conceptualization of PENTA and is expressed explicitly in the presentation of the model. Figure 11r.1 is a reproduction of the schematic of PENTA, now with the addition of optional mappings (indicated by curved arrows) to various underlying levels that are more direct than those assumed in the model. Also added is a representation (the cloud on the far left) of all the meanings that could potentially, but not necessarily, be conveyed by speech. As indicated by the crosses, surface prosody (solid curve on the far right) not only cannot be mapped directly to meanings (longest curved arrow), but also cannot be directly linked to communicative functions, encoding schemes, underlying articulatory targets, or even the target parameters. In fact, at least three degrees of separation were recognized when PENTA was first proposed: articulatory implementation, target assignment, and parallel encoding (Xu 2004a, 2004b). In other words, the very premise of PENTA is that surface "phonetic outcomes" are not mapped directly to meanings. Of course, it is not enough to just point out the mismatches between meaning and phonetic outcomes. PENTA is about how meanings can be ultimately mapped to surface prosody through specific connection mechanisms so that there are no missing conceptual links. This means that each of the three degrees of separation needs to be explicitly represented in the model. Very broadly, as shown in figure 11r.1, meanings are first conventionalized into communicative functions, each having an encoding scheme that has been developed through many rounds of conversational interactions. The encoding schemes of all functions work in parallel to jointly determine a single sequence of targets. These targets are then articulatorily implemented through nonoverlapping sequential target approximation to generate continuous surface acoustic events.

This conceptualization indeed deviates from what Pierrehumbert, in her commentary, calls "modern linguistic theories" of prosody in various ways. In particular, two ideas offered by PENTA, which are mentioned in the main essay of this chapter, are worth recapitulating. The first is that the function-form relation, as formulated by de Saussure (1916), needs a major refinement. The second is that parametric

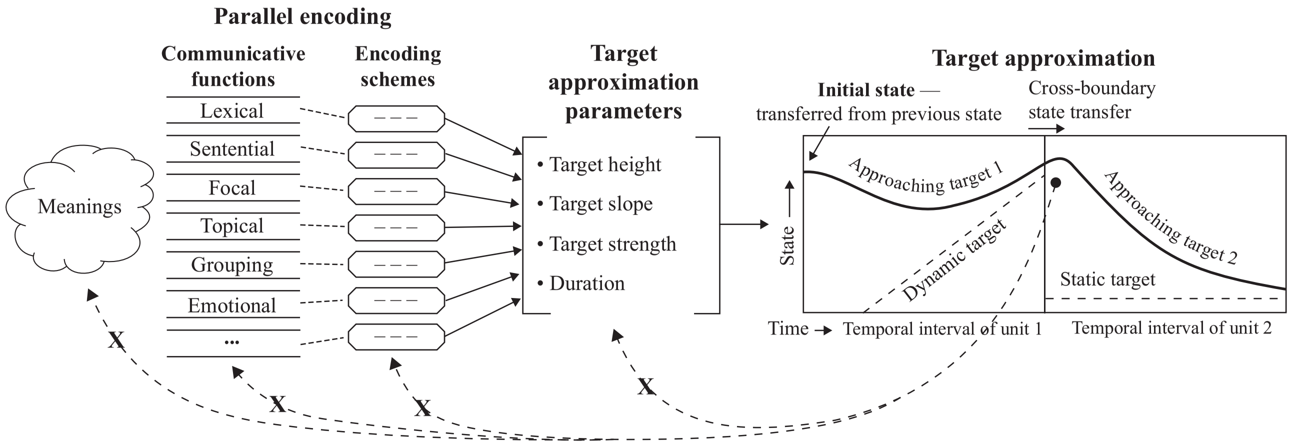

Parallel encoding
Communicative functions
Encoding schemes
Lexical
Sentential
Focal
Topical
Grouping
Emotional
...
Meanings
Target approximation parameters
• Target height
• Target slope
• Target strength
• Duration
Target approximation
Initial state — transferred from previous state
Cross-boundary state transfer
Approaching target 1
Approaching target 2
Dynamic target
Static target
State
Time
Temporal interval of unit 1
Temporal interval of unit 2
X
X
X
X

Figure 11r.1

representations should replace symbolic representations as the final link to surface phonetics. These points are elaborated in this response.

Why Function First?

De Saussure's (1916) notion that linguistic units are unities of signified and signifier does not make it clear what to do if there are uncertainties about both the signifier and the signified. This vagueness has not been a major problem for segmental phonemes because their function is relatively straightforward: to differentiate words. Thanks to people's strong intuition about words, the only major uncertainty is whether a particular segment does or does not distinguish certain words in a particular language. In prosody, both the form of the contrastive units and their functions are often ambiguous, as can be seen in the lack of consensus on both after decades of research. It is thus tempting, and has been tried many times, to first develop a descriptive account of easily observable surface prosodic features such as peaks, valleys, shapes, contours, and overall trends (Bolinger 1986; Crystal 1969; Grabe, Kochanskiand, and Coleman 2007; 't Hart, Collier, and Cohen 1990) with the hope that their meaning associations can be determined by further research. Likewise, units such as pitch accents, phrase accents, and boundary tones were originally summarized from "observed features of F0 contours" without explicit association with meanings, as is made clear in Pierrehumbert (1980, 59). Although there have been later efforts to link them to pragmatic meanings such as truth condition and common ground (Pierrehumbert and Hirschberg 1990), the proposed prosodic units remain primarily defined by their forms, as is evident from the fact that transcription of pitch tracks is used as a major means of prosody analysis (Silverman et al. 1992).

What is overlooked in these approaches is that this is *not* how segmental phonemes are determined. While it is true, as Pierrehumbert notes in her commentary, that "each language has a relatively small inventory of phonological units" ("Introduction"), whether a particular segment should be considered as a phoneme has to be determined by whether it serves to make any specific lexical contrasts rather than whether it sounds sufficiently different from other segments (Swadesh 1934). In other words, a highly specific functional contrast is the primary determinant of the phonemic status of the segment.

What may have made the segmental phonology different from prosody is what is known as *duality of patterning* (Hockett 1960), which is the essence of phonology as a bottleneck that, as Pierrehumbert notes, "helps the language learner to acquire a large vocabulary by allowing articulatory and perceptual patterns exhibited in one word to be reused in other words" ("Introduction"). Here the key word is the *reuse* of the same phoneme in different words, for example, the vowel /i/ in *bin*, *pin*, and *tin*, and the consonants /b/ and /n/ in *bin*, *ban*, and *bun*. Note, however, that the reuse is within the same function, that is, lexical contrast. An appropriate comparison in prosody would be the reuse of on-focus expansion and postfocus compression (PFC) of pitch range in foci at different sentence locations (Xu, Chen, and Wang 2012. But the reuse of the same phonetic feature would not work across functions. It would be hard to claim, for example, that because a postfocus high (H) tone has the same pitch level as a prefocus low (L) tone, the [low] feature is shared between the focus function and the lexical function. In other words, it is unlikely that there is a function-independent phonological /Low/ floating around in its own right, because the [low] is only relative to other tones within the same lexical contrast function.

As recognized by Hockett (1960), duality of patterning is due to heavy crowding in the lexical contrast function, as the number of words that need to be encoded massively exceeds the number of possible distinct segmental categories. Prosody, in contrast, confronts a different kind of crowding, that is, each prosodic dimension, for example, F0, is shared by many functions: lexical, focal, phrasal, topical, sentential, attitudinal, emotional, social-indexical, and so on. To make things worse, the identity and nature of these functions are not clear, given the lack of reference in the form of words, either spoken or written. Faced with this difficulty, PENTA-based research has followed a function-first principle that goes beyond the simple function-form relation envisaged by de Saussure. That is, the task of prosody modeling is to find out whether a particular set of meanings has been conventionalized into a communicative function, and what the encoding scheme of this function is like in terms of how the various prosodic dimensions are utilized to encode its internal categories. Following this principle, observable prosodic forms are always treated as a secondary property, that is, a means of encoding the function-internal categories. This is why PENTA-based studies never use prosodic transcription as a method of prosodic analysis.

Hypothesis Testing by Controlled Experiments

Identifying communicative functions and their encoding schemes is by no means a trivial task. The multiple degrees of separation depicted in figure 11r.1 means that not only are surface acoustic events not directly mapped to meanings, but also no two adjacent levels are linearly related to each other to allow analysis by inversion, that is, deriving the underlying form directly from surface properties. Starting from the right end of figure 11r.1, target approximation, implemented as a generative model in the form of quantitative target approximation (Prom-on, Xu, and Thipakorn 2009), cannot be mathematically inversed to derive the underlying targets. So our modeling work has always used analysis-by-synthesis to estimate the underlying targets (Prom-on, Xu, and Thipakorn 2009; Xu and Prom-on 2014). And even with this approach, the quality of the target estimation is correlated with the size of the training corpus. This means that it is simply impossible to derive authentic underlying targets from single utterances.

Moving leftward to the link between underlying targets and the encoding schemes, any single target is the end result of joint contributions by multiple encoding schemes, which makes it impossible to derive all the contributing encoding schemes from an estimated target, no matter how accurate the estimation may be. Even within an encoding scheme, a large portion of it consists of conventions that stipulate arbitrary context-sensitive assignment of the target parameters (referred to by Pierrehumbert as “language-specific constraints”). For Mandarin, for example, the low tone would assume a rising-tone–like target if it is followed by another low tone. This means that even if a contour is correctly recognized as related to a rising tone, the underlying morpheme could be either one with the low tone or with the rising tone. For English, as found in Liu et al. (2013), whether a stressed syllable is assigned a high or low-rising target depends on its position in word, focus status, and the modality (question or statement) of the sentence. This again means that it is impossible to derive individual functions even from the estimated targets.

Finally, as indicated at the far left of the figure, not all possible meanings have conventionalized functions. It is therefore impossible to know, a priori, whether a potential meaning, no matter how useful it may seem (e.g., truth condition and common ground), can be mapped to a specific encoding scheme. For example, seven different types of focus have been suggested in Gussenhoven (2007). But so far, not even the two most obviously different types, namely, information focus and contrastive focus, have been demonstrated to be consistently distinct from each other in their prosodic realizations (Hanssen, Peters, and Gussenhoven 2008; Hwang 2012; Katz and Selkirk 2011; Kügler and Ganzel 2014; Sityaev and House 2003).

In the face of so many levels of indirect and nonunique mappings, the only viable method of discovering whether a potential meaning has developed a conventionalized function, and what the encoding scheme of that function is like, is hypothesis testing by controlled experiments. In this paradigm, both the function and the encoding schemes are treated as hypothetical, and experiments designed to systematically manipulate the functional content are performed. In the end, it is the outcome of the experiments, which often requires multiple studies, that can inform us, with various levels of certainty, of the presence of a function and the internal structure of its encoding scheme. It is with this approach, for example, that it is determined that the most salient encoding feature of prosodic focus is PFC of pitch range and intensity in many languages and that PFC is nevertheless fully absent in many other languages (Xu, Chen, and Wang 2012).

Even with controlled experiments, however, there is an issue of whether function- or form-defined units should be the target of testing. For example, when pitch accent is targeted in some controlled studies (e.g., Grabe et al. 2000; Shue et al. 2010; Turk and White 1999), the method of elicitation is the same as those used in studies of focus, that is, question-answer or negation paradigms (Cooper, Eady, and Mueller 1985; Eady and Cooper 1986; Liu et al. 2013; Patil et al. 2008; Wang and Xu 2011; Xu and Xu 2005). Due to the presumption of pitch accents as phonological units, these studies either examine phonetic properties of the focused words only or treat those of postfocus components as due to phrase accent or boundary tones that are independent of the nuclear pitch accents.

From the perspective of the function-first principle, pitch accents are merely a phonetic property, as they are identified by the presence of local F0 peaks, valleys, or movements that sound and/or look prominent, which may or may not be due to focus. For example, a prominent F0 peak may occur at the beginning of an utterance even in the absence of an initial focus (Wang and Xu 2011). Or, a prominent pitch movement may occur near the end of a sentence, which would, by definition, be treated as a nuclear pitch accent. But both production and perception studies have shown that these peaks would neither be always intended nor perceived as a sentence-final focus (Cooper, Eady, and Mueller 1985; Rump and Collier 1996; Xu and Xu 2005). Furthermore, focus may not always be marked by an F0 peak more prominent than that in a neutral-focus sentence, as found in Turkish (Ipek 2011). This is not surprising, because the presence of PFC (which is attributed to deaccenting and/or an L-phrase accent in the autosegmental-metrical [AM] theory) already enables successful perception of focus (Ipek 2011; Rump and Collier 1996; Xu, Xu, and Sun 2004). Focus, therefore, is empirically attested as a communicative function marked by multiple phonetic cues, including on-focus increase of pitch range, intensity, and duration, and postfocus reduction of pitch range and intensity (Xu 2011), with a temporal domain that expands even across a silent phrasal pause within a sentence (Wang, Xu, and Ding 2018). In contrast, pitch accent, even when seemingly obvious, is only one of such cues, which may not even be the most critical cue, because the presence of an F0 peak later in the utterance would effectively block the perception of an early focus (Rump and Collier 1996). It would therefore be difficult for PENTA to equate focus with nuclear accent in the phrase, as suggested in Pierrehumbert's commentary.

By the same token, boundary tone, as a cue to sentence modality (question versus statement), is also only one of the phonetic markers of the contrast, rather than being a phonological unit in its own right. For American English, at least, the marking of modality involves not only a sentence-final F0 rise or fall, but also a drastic raising or lowering of postfocus F0 register (treated as due to an independent phrase accent in the AM theory), and a change of target height and target slope of all stressed syllables throughout the sentence (Liu et al. 2013).

Economy of Representation and Degrees of Freedom

The kind of controlled experiments involved in typical empirical studies, however, can go only so far as identifying the functions and the gross patterns of their encoding schemes. To be able to account for the full details of surface prosody, a further step is needed to establish a form of representation that can generate real speech–like continuous prosodic events. This ultimate goal is attempted in PENTA through parametric representation. In this regard, however, PENTA is often criticized for being uneconomical in representation (see Arvaniti, chapter 1, this volume; Arvaniti and Ladd 2009, 2015), given its insistence on (i) pitch target for every syllable even if it is unstressed or bearing the neutral tone, and (ii) full specification of all targets in terms of not only target height (register), but also target slope and target strength, with no allowance for any underspecifications. But we fully agree with Pierrehumbert's remark that "the human cognitive system can learn very detailed patterns and often represents them with a great deal of redundancy" ("Conclusion"). The redundancy is not only in terms of the multiple cues for any specific communicative function, as we've discussed, but also in terms of detailed continuous trajectories that carry massive variability due to articulatory mechanisms, dialectical differences, and idiosyncrasies of individual speakers.

The solution to the redundancy problem explored in the PENTA approach, as detailed in the main essay of this chapter, is model-based parametric representation. *Model-based* means that the representation is meaningful only with respect to a specific computational model. *Parametric* means that targets are specified by numerical parameters rather than symbolic features. The representation of F0, for example, is by numerical specifications of target height, target slope, and target strength, as shown in figure 11r.1. The parameter values are obtained neither by transcription nor by direct acoustic measurement, but by training the computational model on real speech data. Depending on the nature of the training data, the learned targets can be language-, dialect-, or speaker-specific. Our computational studies so far have shown that the approach is able to generate pitch contours that are both natural sounding and functionally contrastive (Prom-on, Thipakorn, and Xu 2009; Xu and Prom-on 2014). And our pilot results based on speech corpora that are less well controlled than typical experimental data have also been encouraging.

Overall, whether a representation is sufficiently economical cannot be measured by the number of representational units assumed by a theory, but by the total specifications needed to generate detailed continuous prosodic events that resemble those of natural speech. If a unit is specified only in terms of H or L, as is the case with pitch accents, phrase accent, and boundary tones, somewhere down the line, there have to be specifications of the exact pitch height, the onset time and offset time of the unit, and how exactly the unit is connected to adjacent units. If underspecification is assumed, sooner or later there has to be a mechanism to generate surface acoustics for the underspecified units. Without including all these specifications, it is impossible to compare degrees of freedom between different models.

Another way of assessing the economy of a model is to see how many redundant parameters are required. PENTA uses only three free parameters: height, slope, and strength of targets. None of them is redundant, because they are all independently motivated. *Target height* is motivated by its universal recognition; *target slope* is motivated by the consistency of final velocity in dynamic tones (Wong 2006; Xu 1998); and *target strength* is motivated by the sluggish realization of a mid target in the neutral tone in Mandarin (Chen and Xu 2006) and unstressed syllable in English (Xu and Xu 2005). In comparison, the equivalent of target strength in the Fujisaki and the task dynamic models (stiffness) is mostly fixed (Fujisaki 1983; Saltzman and Munhall 1998) and so is largely redundant. On the other hand, the temporal domain of target approximation is fixed to the entire syllable in PENTA (Xu and Prom-on 2015), so that there are virtually no temporal degrees of freedom. This also contrasts with the Fujisaki model (Fujisaki 1983) and articulatory phonology/task dynamic model (Browman and Goldstein 1992; Saltzman and Munhall 1989), where the onset and offset of the commands and gestural scores are free parameters, which means many more degrees of freedom in the temporal domain than PENTA. Given that the AM theory has no strict specifications of tonal alignment, it would also face the problem of degrees of freedom in the temporal domain.

Conclusion

PENTA is part of an effort to develop a new way of conceptualizing the mapping between meanings and continuous acoustic signals in speech, starting from the prosodic aspect. The multifold complexity of prosody has forced us to go back to the first principles to reconsider the phonetic-phonology interface in light of the function-form dichotomy. As a result, PENTA is one of the most indirect models of prosody, as it explicates multiple degrees of separation between meaning and continuous surface prosody. At the same time, it also insists that there be no broken links in the theoretical conceptualization of prosody and intonation and has implemented this tenet by proposing specific connection mechanisms in its computational implementation. What has also emerged from this effort is that model-based parametric representation could be the key to understanding not only the mapping of meaning to continuous phonetic output, but also how the acquisition of speech production is achieved (Xu and Prom-on 2014, 2015).

References

Arvaniti, A., and D. R. Ladd. 2009. “Greek wh-Questions and the Phonology of Intonation.” *Phonology* 26 (1): 43–74.

Arvaniti, A., and D. R. Ladd. 2015. “Underspecification in Intonation Revisited: A Reply to Xu, Lee, Prom-on and Liu.” *Phonology* 32:537–541.

Bolinger, D. 1986. *Intonation and Its Parts: Melody in Spoken English*. Palo Alto: Stanford University Press.

Browman, C. P., and L. Goldstein. 1992. “Articulatory Phonology: An Overview.” *Phonetica* 49:155–180.

Chen, Y., and Y. Xu. 2006. “Production of Weak Elements in Speech: Evidence from f0 Patterns of Neutral Tone in Standard Chinese.” *Phonetica* 63:47–75.

Cooper, W. E., S. J. Eady, and P. R. Mueller. 1985. “Acoustical Aspects of Contrastive Stress in Question-Answer Contexts.” *Journal of the Acoustical Society of America* 77:2142–2156.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

de Saussure, F. 1916. “Nature of the Linguistics Sign.” In *Cours de linguistique générale*, edited by C. Bally and A. Sechehaye, 66–70. New York: McGraw-Hill.

Eady, S. J., and W. E. Cooper. 1986. “Speech Intonation and Focus Location in Matched Statements and Questions.” *Journal of the Acoustical Society of America* 80:402–416.

Fujisaki, H. 1983. “Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing.” In *The Production of Speech*, edited by P. F. MacNeilage, 39–55. New York: Springer-Verlag.

Grabe, E., G. Kochanski, and J. Coleman. 2007. “Connecting Intonation Labels to Mathematical Descriptions of Fundamental Frequency.” *Language and Speech* 50:281–310.

Grabe, E., B. Post, F. Nolan, and K. Farrar. 2000. “Pitch Accent Realization in Four Varieties of British English.” *Journal of Phonetics* 28:161–185.

Gussenhoven, C. 2007. “Types of Focus in English.” In *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, edited by C. Lee, M. Gordon and D. Büring, 83–100. New York: Springer.

Hanssen, J., J. Peters, and C. Gussenhoven. 2008. “Prosodic Effects of Focus in Dutch Declaratives.” In *Proceedings of Speech Prosody*, edited by P. A. Barbosa, S. Madureira, and C. Reis, 609–612.

Hockett, C. F. 1960. “The Origin of Speech.” *Scientific American* 203:88–96.

Hwang, H. K. 2012. “Asymmetries between Production, Perception and Comprehension of Focus Types in Japanese.” In *Proceedings of Speech Prosody 2012*, edited by Q. Ma, H. Ding, and D. Hirst, 326–329.

Ipek, C. 2011. “Phonetic Realization of Focus with no On-Focus Pitch Range Expansion in Turkish.” In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by Wai-Sum Lee and Eric Zee, 140–143.

Katz, J., and E. Selkirk. 2011. “Contrastive Focus vs. Discourse-New: Evidence from Phonetic Prominence in English.” *Language* 87 (4): 771–816.

Kügler, F., and S. Genzel. 2014. On the Elicitation of Focus: Prosodic Differences as a Function of Sentence Mode of the Context? *Proceedings of the 4th International Symposium on Tonal Aspects of Languages*, edited by C. Gussenhoven, Y. Chen, and D. Dediu, 71–74.

Liu, F., Y. Xu, S. Prom-on, and A. C. L. Yu. 2013. “Morpheme-Like Prosodic Functions: Evidence from Acoustic Analysis and Computational Modeling.” *Journal of Speech Sciences* 3 (1): 85–140.

Patil, U., G. Kentner, A. Gollrad, F. Kügler, C. Féry, and S. Vasishth. 2008. “Focus, Word Order and Intonation in Hindi.” *Journal of South Asian Linguistics* 1:55–72.

Pierrehumbert, J. 1980. “The Phonology and Phonetics of English Intonation.” PhD diss., MIT.

Pierrehumbert, J., and J. Hirschberg. 1990. “The Meaning of Intonational Contours in the Interpretation of Discourse.” In *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack, 271–311. Cambridge, MA: MIT Press.

Prom-on, Y., S. Xu, and B. Thipakorn. 2009. “Modeling Tone and Intonation in Mandarin and English as a Process of Target Approximation.” *Journal of the Acoustical Society of America* 125:405–424.

Rump, H. H., and R. Collier. 1996. “Focus Conditions and the Prominence of Pitch-Accented Syllables.” *Language and Speech* 39:1–17.

Saltzman, E. L., and K. G. Munhall. 1989. “A Dynamical Approach to Gestural Patterning in Speech Production.” *Ecological Psychology* 1:333–382.

Shue, Y.-L., S. Shattuck-Hufnagel, M. Iseli, S.-A. Jun, N. Veilleux, and A. Alwan. 2010. “On the Acoustic Correlates of High and Low Nuclear Pitch Accents in American English.” *Speech Communication* 52 (2): 106–122.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. “ToBI: A Standard for Labeling English Prosody.” In *Proceedings of the 1992 International Conference on Spoken Language Processing*, edited by J. J. Ohala, T. Nearey, B. Derwing, M. Hodge, and G. Wiebe, 867–870.

Sityaev, D., and J. House. 2003. “Phonetic and Phonological Correlates of Broad, Narrow and Contrastive Focus in English.” In *Proceedings of the Fifteenth International Congress of Phonetic Sciences*, edited by D. Recasens, M.-J. Solé, 1819–1822.

Swadesh, M. 1934. “The Phonemic Principle.” *Language* 10:117–129.

’t Hart, J., R. Collier, and A. Cohen. 1990. *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Turk, A. E., and L. White. 1999. “Structural Influences on Accentual Lengthening.” *Journal of Phonetics* 27:171–206.

Wang, B., and Y. Xu. 2011. “Differential Prosodic Encoding of Topic and Focus in Sentence-Initial Position in Mandarin Chinese.” *Journal of Phonetics* 39 (4): 595–611.

Wang, B., Y. Xu, and Q. Ding. 2018. “Interactive Prosodic Marking of Focus, Boundary and Newness in Mandarin.” *Phonetica* 75 (1): 24–56.

Wong, Y. W. 2006. “Realization of Cantonese Rising Tones under Different Speaking Rates.” In *Proceedings of Speech Prosody 2006*, edited by R. Hoffmann and H. Mixdorff, PS3-14-198.

Xu, Y. 1998. “Consistency of Tone-Syllable Alignment across Different Syllable Structures and Speaking Rates.” *Phonetica* 55:179–203.

Xu, Y. 2004a. “The PENTA Model of Speech Melody: Transmitting multiple Communicative Functions in Parallel.” In *Proceedings of From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, edited by J. Slifka, S. Manuel and M. Matthies, C-91–96.

Xu, Y. 2004b. “Transmitting Tone and Intonation Simultaneously: The Parallel Encoding and Target Approximation (PENTA) Model.” In *Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, edited by B. Bel and I. Marlien, 215–220.

Xu, Y. 2011. “Post-Focus Compression: Cross-Linguistic Distribution and Historical Origin.” In *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, edited by Wai-Sum Lee and Eric Zee, 152–155.

Xu, Y., S.-W. Chen, and B. Wang, B. 2012. “Prosodic Focus with and without Post-Focus Compression (PFC): A Typological Divide within the Same Language Family?” *Linguistic Review* 29:131–147.

Xu, Y., and S. Prom-on. 2014. “Toward Invariant Functional Representations of Variable Surface Fundamental Frequency Contours: Synthesizing Speech Melody via Model-Based Stochastic Learning.” *Speech Communication* 57:181–208.

Xu, Y., and S. Prom-on. 2015. “Degrees of Freedom in Prosody Modeling.” In *Speech Prosody in Speech Synthesis—Modeling, Realizing, Converting Prosody for High Quality and Flexible speech Synthesis*, edited by K. Hirose and J. Tao, 19–34. Berlin: Springer.

Xu, Y., and C. Xu. 2005. “Phonetic Realization of Focus in English Declarative Intonation.” *Journal of Phonetics* 33 (2): 159–197.

Xu, Y., C. X. Xu, and X. Sun. 2004. “On the Temporal Domain of Focus.” In *Proceedings of International Conference on Speech Prosody*, edited by K. Hirose, 81–84.

Index

Abercrombie, D., 121, 122, 225, 341
Abramson, Arthur, 20n8
Accent commands, 4
Adamou, Evangelia, 39
Adams, J. *See* Glewwe et al. (2018), 15
African languages and linguistics, 46, 186, 187, 415
African tone languages, 29, 142, 153, 158, 172n4, 416, 417
Agarwal, A. *See* Shouval et al. (2014), 242
Aguilar, Lourdes. *See* Escudero et al. (2012), 71
Ahn, B., 165, 210
Alignment, and tonal targets, 29
  and phonetic implementation in AM model, 40–41
Alt tier, 162
AM. *See* Autosegmental-metrical (AM) approach to prosody
AM Tones and Break Indices (G-ToBI) system German, 3
Ambrazaitis, G. I., 288, 302, 303
Analysis-by-synthesis, 11–14
Analytic bias, 14
Anchor normalization, 356–357
Anderson, A., 165
Anderson, Poul, 85, 87–88, 89
Annotation, prosodic, 117, 145
  emphasis, 124
  F0 transitions, 130–133
  functions, prosodic, 119–120
  levels of representation, 117–119
  macromelody, 127–130
  micromelody, 127–130
  Momel representation, 133–136
  phonetic representation, 125–136
  and ProZed, 141
  rhythm units, 121, 124
  stress foot, 121
  structure, prosodic, 120–124
  surface phonological representation, 136–141
  terminality, 124
  tonal units, 121–124
  underlying phonological representation, 142–144
Arabic languages, 31, 32, 69, 130, 139, 417
Arai, T., 409
Arbisi-Kelm, T., 19, 161
Armstrong, L., 117, 142
Armstrong, Megan E., 48, 165, 169
Arnhold, Anja, 37, 51, 163
Arnholtz, A., 85, 86, 89, 95
Arnold, D., 293
Arnold, G. F., 46, 117, 142, 184, 290, 324, 331–336, 340
Aronoff, M. *See* Meir et al. (2013), 415; Sandler et al. (2011), 8
Articulatory phonology, 218–219
  task dynamics (AP/TD), 237–243
Arvaniti, Amalia, 26, 27, 29, 30, 33–34, 35, 39, 40, 41, 42, 43, 44–45, 46, 47, 49, 50, 51, 52, 64, 65, 68, 69, 70, 71, 77, 78, 79–81, 104, 156, 157, 163, 167, 182, 187, 192, 194, 196n14, 204, 208, 225, 248, 251, 301, 339, 416, 422, 430
  See also del Giudice et al. (2007), 183; Grice et al. (2000), 7; Prom-On et al. (2016), 34
Ashby, Michael, 259
Ashby, Patricia, 259
Aston, C. H., 225
Asu, E. L., 288
Atterer, Michaela, 40, 164
Auchlin, A. 272
  See also Avanzi et al. (2010), 278fig7.6; Goldman, Auchlin et al. (2007), 272
Auer, P., 308
Auran, C., 141, 145n3

Australian National Database of Spoken Language (ANDOSL), 165
Autosegmental-metrical (AM) approach to prosody, 3–4, 25–28
  absence of representational uniformity, 7
  and AM models of English intonation, 204–205
  AM Tones and Break Indices (G-ToBI) system (German), 3
  applications of, 50–52
  and composition of melodies, 35–37
  and declination, 42
  descriptive inadequacy and inconsistency in, 184–185
  distinction between phonology of intonation
  and downstep, 34–35
  enduring insights from, 183–184
  enhanced AM, 182–183
  essentials of, 28–29
  events, 186
  fundamental frequency (F0) contours in
  importance of intonation in, 25–26
  inconsistencies in mapping pitch accents to F0
  and intonational meaning, 46–48
  and intonational typology, 48–50, 79–81
  main advantage of, 64–68
  models of prosody, 3, 4–5
  and phonetic implementation, 39–45
  phonetic interpolations versus tones, 186–187
  and phonetic realization, 25–26
  and phonetic variability, 77–79
  pitch level and pitch span, 41
  problems with phonetic rules, 185–186
  and prosodic typology, 76–77
  and the RaP transcription system, 188–195, 207–210
  and scaling, 34–35
  and secondary association, 32–34
  and structure of tonal events, 29–32
  syntagmatic tonal relationships in, 206–207
Autosegmental string, and metrical structure, 37–39
Autosegments, 28
Avanzi, M., 272, 278fig7.6
  See also Delais et al. (2015), 152; Goldman, Auchlin, et al. (2007), 272
Avesani, Cinzia, 39, 69
  See also Grice et al. (2005), 165, 168, 169
Aylett, Matthew, 71, 151, 251, 368
Ayoker, Otto Gwado, 40
Ayres Elam, G. M., 151, 207, 247

Bader, M. *See* Anderson et al. (1991), 165
Baek, S., 163, 170
Bailly, G., 385
Baken, R. J., 262
Baker, W., 164
Baltazani, Mary, 25, 26, 27fig1.1, 35, 39, 40, 42, 43, 47, 48, 51, 157, 208, 248
Bambara language, 124
Bangalore, S., 369
Barbisch, M., 358
Barbosa, P. A., 226, 238
Bardiaux, A., 272
Barnes, Jonathan, 17, 34, 43, 44, 72, 182, 183, 199, 266, 311, 313
  See also Brugos et al. (2018), 19
Barry, W. J., 301
Bartolo, R. *See* Merchant et al. (2008), 242
Basbøll, Hans, 85, 89, 111
Basque language, 29
Batliner, Anton. *See* Grice et al. (1996), 248, 251, 368
Baudouin de Courtenay, Jan, 20n7
Baumann, Stefan, 31, 32, 35, 69, 156, 251, 298, 353, 359, 361, 367
  See also Cangemi et al. (2016), 72; Kügler et al. (2015), 295, 308
Bayerl, P. S., 368
Beaugendre, F., 263, 270
Beaver, David, 39, 46, 412, 416
  See also Calhoun et al. (2010), 353, 365
Beckman, Mary, 25, 28, 31, 32–33, 34, 35, 37, 38, 39, 42, 44, 49, 50–51, 52, 53n5, 65, 71, 76, 120, 151, 152, 154, 155, 156, 157, 158, 159, 163, 166, 172n4, 173n13, 174n22, 182, 185, 186, 188, 192, 195n9, 196n14, 205, 206, 207, 213, 219, 242, 247, 249, 251, 252, 338, 353, 368, 395, 408, 409, 411, 414, 418, 419, 420
  See also Peng et al. (2005), 156, 159, 166; Silverman et al. (1992), 50, 117, 136, 137, 151, 247, 338, 414, 418, 427; Syrdal et al. (2001), 368
Beddor, Patrice Speeter, 227, 417
Bell, C. C., 11, 12, 13
Bengali language, 37, 51, 152, 166, 168, 173n12, 173n17, 415
Ben Jannet, M. *See* Loevenbruck et al. (2013), 292

Benzmüller, Ralf, 31, 32, 35, 151, 156, 359, 361, 367
  See also Grice et al. (1996), 248, 251, 368
Berber language, 40, 46, 65, 66, 67, 68, 76, 77
Berent, I., 195n6
Bergmann, A., 165
Berkovits, R., 218
Bernstein, N. A., 382
Bertinetto, P. M.,139
Best, Catherine T.
Beutnagel, M. *See* Syrdal et al. (2000), 368
Bias (cognitive predisposition)
  analytic, 14
  channel, 14
  substantive, 15
Bigi, B., 141
Billon, M., 241
Birch, Stacy, 247, 412
Bird, C. S., 124
Birkholz, P. *See* Xu et al. (2013), 401
Bishop, Jason, 69, 156, 163, 165, 166, 170, 183, 209
Bitonal accents, 30
Black, A. W. *See* Taylor et al. (1999), 357
Blasi, D. *See* Dingemanse et al. (2015), 415
Bloch, B., 289
Bloch, S. N. J., 86
Bo, Alf, 86, 99
Boersma, P., 108, 133–134, 263, 291
Bolinger, Dwight, 46, 47, 123–124, 153, 206, 225, 239, 251, 252, 273, 274, 276, 333, 341, 385, 408, 427
Bolivar, M. *See* Dilley et al. (2006), 171, 275
Bombien, L., 216, 218, 220
Bomhard, A. R., 399
Bootsma, R. J., 241
Borràs-Comes, Joan, 42, 48
Borys, S. *See* Hasegawa-Johnson et al. (2005), 369
Boston University Radio Speech Corpus, 36, 165, 412
Boundary tones, 28
Bouzon, C., 124, 145n3
Boves, 81n1
Boyce, Suzanne, 42
Boyle, E. *See* Anderson et al. (1991), 165
Brainard, M. S., 396
Brand, T., 356
Branderud, P
Braun, Bettina, 69
  See also Kügler et al. (2015), 295, 308
Break indices tier, 160–161
Breen, M., 171, 189, 192
  See also Brugos et al. (2008), 162; Dilley et al. (2006), 171, 275
Bregman, A.S., 188
Brenier, J. *See* Calhoun et al. (2010), 353, 365
Bresch, E. *See* Byrd et al. (2009), 215
Brink, L., 85
British school of intonation, 36, 47, 319, 344–346
  and AM, 338–344
  heyday of, 329–338
  history of, 319–329
Brognaux, Sandrine, 280
Browman, Catherine P., 29, 213, 214, 215, 227n1
Brown, G., 166, 329
Brown, Meredith, 40, 171, 184, 185–186, 189, 195n9, 196n14, 367
  See also Breen et al. (2018), 194
Bruce, Gösta, 32, 44, 90, 98, 152, 290, 367, 417, 420
Bruggeman, Anna, 66, 67, 68
Brugos, Alejna, 19, 50–51, 162, 247
  See also Barnes et al. (2010), 34; Barnes et al. (2012), 43, 72, 182, 183, 189; Barnes et al. (2013), 43; Barnes et al. (2014), 266
Brunner, J., 215
Bryant, Gregory A., 46, 292
Buckeye Corpus, 165
Bühler, K., 305
Bulgarian language, 17
Bunnell, H., Timothy, 38
Bunz, H., 214
Burdin, Rachel, 71
Büring, Daniel, 46, 353
Burnett, James, 322, 323
Burnett, T. A., 189
Burns, E. M., 184, 188
Buthke, C. *See* Delais et al. (2015), 152
Butler, C., 322, 346n1
Byrd, Dani, 38213, 214, 215, 216, 217–218, 219–220, 223, 224, 225, 227n1, 237, 242
  See also Goldstein et al. (2009), 224; Nam et al. (2004), 215, 221
Byrne, W. *See* Ostendorf et al. (2001), 165

Caley, R. *See* Taylor et al. (1999), 357
Calhoun, Sasha, 36, 39, 166, 353, 365, 366
  See also Schweitzer et al. (2015), 366
Calleja, Nagore, 40

Calling contour, 193
Cambier-Langeveld, Tina, 38, 218
Campbell, Nick, 39, 136
Campbell, W. N., 238
Campione, E., 134table3.1
Cangemi, Francesco, 71, 72, 167
  See also Bruggeman et al. (2018), 72; Niebuhr et al. (2011), 299, 312
Cantonese language, 51, 399
Carletta, J. *See* Calhoun et al. (2010), 353, 365
Carlson, K., 219
Carlson, R., 288
Carmichael, L. *See* Ostendorf et al. (2001), 165
Carvey, H., 192
Casserly, E., 165, 219
Catalan language, 78, 79, 138, 152, 169
  bitonal accents in, 31
  multitonal combinations of edge tones in, 31
  scaling in, 42
  and secondary association of phrase accents, 34
  tone gestures in, 222
Catford, J. C., 20n10, 20n11
Chahal, Dana, 31, 32, 39
Champagne-Lavau, M. *See* Loevenbruck et al. (2013), 292
Chan, Marjorie K. M., 51, 156, 157, 159, 173n13
Chang, Y.-C., 196n17
Channel bias, 14
Chavarría, S. *See* Yoon et al. (2004), 368, 369
Chen, Aoju, 46, 292, 353
Chen, K. *See* Hasegawa-Johnson et al. (2005), 369
Chen, L. *See* Goldstein et al. (2009), 224
Chen, M. Y., 398
Chen, S., 189
Chen, S.-W., 399, 400, 427, 429
Chen, Ying, 195n8, 231, 380, 382, 384, 399
Cheney, D. L., 8
Cheng, C., 380
Chentir, A., 130
Chickasaw language, 37, 40, 151, 157, 158, 159, 160, 161, 162
Chinese, Standard (Beijing), 139, 156, 391
Chitoran, I. *See* Goldstein et al. (2009), 224
Cho, H., 136
Cho, Taehong, 38
  See also Keating et al. (2004), 216
Choi, J.-Y. *See* Hasegawa-Johnson et al. (2005), 369
Chomsky, Noam, 6, 15, 119, 323, 409
Christiansen, M. H. *See* Dingemanse et al. (2015), 415
Christodoulides, G., 272
Chuenwattanapranithi, S., 388, 401
Cieri, C., 388
Clark, Brady Zack. *See* Beaver et al. (2007), 39, 46, 412, 416
Classe, A., 238
Clements, G. N., 7, 42, 142, 191
Clifton, Charles, Jr., 219, 247, 412
Clitic Group, 38
Cobler, M., 191
Coetzee, A., 195n6
Cognition, and prosodic models, 15–16
Cohen, Antonie, 40, 42, 46, 78, 125, 196n14, 262–263, 264, 266, 271, 273, 275, 427
  See also Hasegawa-Johnson et al. (2005), 369
Co-intrinsic micro-prosody, 265
Cole, Jennifer, 43, 163, 167, 170, 174n21, 184, 225, 226, 238
  See also Hasegawa-Johnson et al. (2005), 369; Yoon et al. (2004), 368, 369
Coleman, J., 328, 420, 427
Collier, René, 40, 42, 46, 78, 98, 125, 262, 263, 264, 266, 273, 275, 427, 429
Commands, accent and phrase, 4
Competence, linguistic, 15
Composition of melodies, 35–37
Configurations, 40
Conkie, A., 351, 358
  See also Syrdal et al. (2000), 368
Connell, Bruce, 42, 384
Context, and intonational meaning, 46
Contrast theory of intonation perception (C-TIP), 311
Conway, E., 251
Cooper, W. E., 154, 220, 362, 385, 399, 429
  See also Eady et al. (1986), 399
Corcos, D. M., 242
Couper-Kuhlen, E., 117, 308
Cowie, R., 396
Cruttenden, Alan, 251, 275, 321, 322, 325, 326, 336–338, 339
Cruz, Marisa. *See* Frota et al. (2015), 152
Crystal, D., 117, 153, 275, 322, 323, 335, 408, 427
Cummins, F., 226
Currie, K., 166

Cutler, A., 184
Cypriot language, 33, 38, 46

Dahan, D., 184
Dainora, Audra, 36, 47, 49
Daland, R. *See* Glewwe et al. (2018), 15
d'Alessandro, C., 189, 263–264, 267, 270
Dalla Bella, S. *See* Pfordresher et al. (2015), 189, 196n17
Dalton, Martha, 40
Daly, Nicola, 41, 52n1
Danish language, intonation in, 85, 86–87, 111–112, 167
  core linguistic function of intonation, 87
  cues to modality, 100
  default sentence accent, lack of, 90
  descriptions, previous, 85–86
  emphasis for contrast by reduction, 90–92
  empirical investigations of, 87–88
  extended model of, 102–103
  F0 pattern in stress group, 89
  final boundary tone, lack of, 92
  final lengthening, absence of, 92–94
  final rise in questions, lack of, 94–95
  focus by reduction, 90
  "high" in questions, lack of, 94–95
  intonation contours, utterance, 98–100
  longer utterances and prosodic boundaries, 100–102
  and meaning, 107–110
  methodology, 88
  modality, cues to, 100
  paucity of prosodic components in Danish, 88–89
  prominence, equal and unequal, 89–92
  and psychological reality, 110
  short utterances, modeling, 98
  speech, read and spontaneous, 110–111
  stressed syllables carry intonation contours, 95–98
  stress group, prosodic, 89
  subordination and superposition, 104–107
  and syntactic structure, 100
  and transcription, 108–110
Dankovica, J., 409
Dauer, R. M., 226, 238
Davidson, K. *See* del Giudice et al. (2007), 183
Dechongkit, S., 384
Declination, in AM model, 42
de Lacy, P., 192, 195n3
Delais-Roussarie, E., 152
del Giudice, Q., 183
De Looze, C., 141, 262, 272
de Moraes, J. A., 272
Demorest, S. M. *See* Pfordresher et al. (2015), 189, 196n17
den Os, E., 353
de Pijper, J. R., 219
Derksen, A. G. *See* Spaai et al. (1993), 262
Dermody, P. *See* Miller et al. (1992), 165
de Ruiter, J., 353
de Saussure, F., 8, 20n7, 314, 385, 407, 425, 427, 428
Diachronic explanations of cross-language patterns, 14–15
Díaz-Campos, M. *See* Beckman et al. (2002), 152
Di Cristo, Albert, 51, 117, 118, 125, 134, 137, 143, 145n6, 163, 196n13, 207, 262, 275
  See also Delais et al. (2015), 152; Rossi et al. (1981), 263, 275
Diehl, R. L., 184
Dilley, Laura C., 20n6, 30, 38, 40, 165, 171, 182, 184, 185–186, 187, 188, 189, 191, 275
  See also Breen et al. (2018), 194; Breen et al. (2012), 171, 189, 194, 196n11; Breen et al. (2014), 194; Miller et al. (1992), 165
D'Imperio, Mariapaola, 34, 37, 40, 43–44, 48, 183, 187, 189, 223
  See also Grice et al. (2005), 165, 168, 169; Loevenbruck et al. (2013), 292; Niebuhr et al. (2011), 299, 312
Dingemanse, M., 415
Diphones, 13–14, 20n6
Dobroth, K. M., 165
Dogil, G., 357, 364
  See also Barbisch et al. (2007), 358; Schweitzer et al. (2015), 366; Wade et al. (2010), 366
Doherty, G. *See* Anderson et al. (1991), 165
Douglas-Cowier, E., 396
Doupe, A. J., 396
Dombrowski, E., 287, 292, 301
Dowling, W. J., 184, 188, 196n17
Downstep, and AM model, 29, 34–35
Duality of patterning, 410, 427
Dupoux, E., 164
Dutch language, 31, 32, 33, 34, 36, 38, 40, 41, 42, 164, 217, 218, 247, 251, 290, 367
Dyhr, N. J., 110–111

Eady, S. J., 399, 429
Eckart, K., 353
Edge tones, 28, 76
Edlund, J., 291
Edwards, Jan, 25, 39, 174n22, 242
Egyptian Arabic language, 32
Eide, E. M., 419
Elordieta, Gorka, 28, 40
  See also D'Imperio et al. (2005), 37
El Zarka, Dina. *See* Bruggeman et al. (2018), 72; Cangemi et al. (2016), 72
Emphasis, 124, 305–306, 323
Engdahl, E., 412
Engelkamp, J., 288
English Broadcast News Speech, 165
English language, 249, 250
  bitonal accents in, 30, 31
  edge tones in, 28
  intonational encoding of focus in, 48–49
  intonational grammar of, *37*
  intonational systems in, 46–48
  melodies in AM analysis, 26–27, 32
  number of levels in the prosodic hierarchy, 37
  nuclear contour units in, 36
  phonetic scaling for, 35
  secondary association of phrase accents in, 32, 34
  uptalk, across varieties of, 46
Equivalent regular bandwidth (ERB), 52n1
Eriksson, A., 238
Escudero, David, 71
Espesser, Robert, 51, 117, 118, 134, 262–263
  See also D'Imperio et al. (2007), 223
Esteve-Gibert, N., 227
Ewan, W. G., 221
Exemplar theory, 365–367

Face, T., 152
Fant, Gunnar, 11, 184, 239, 293
Farrar, Kimberley, 71, 343
  See also Grabe et al. (2000), 40, 44, 429
Faulkner, Dan. *See* Ladd et al. (1999), 78
Faulkner, Hanneke. *See* Ladd et al. (1999), 78
Faure, G., 117, 145n1
Fay, R., 188, 196n17
Fedorenko, E., 195n3
  See also Breen et al. (2010), 193
Feldhausen, I. *See* Delais et al. (2015), 152
Féry, Caroline, 46, 65, 69, 193, 353, 361, 367
Final lowering, 42
Fischer-Jørgensen, Eli, 86, 89, 95
Fiscus, J., 165
Fisher, W. *See* Fiscus et al. (1998), 165
Fitzroy, 192
Flemming, Edward. *See* Beaver et al. (2007), 39, 46, 412, 416
Fletcher, Janet, 30, 32, 40, 46, 50, 51, 77, 156, 165, 169, 242, 420
  See also Stirling et al. (2001), 165
Fodor, J. D., 165, 173n14
Fogerty, D., 194
Foltz, A., 165, 166
Fong, C. *See* Price et al. (1991), 154, 247
Forchhammer, G., 85, 86
Ford, K., 142
Fougeron, Cécile, 38, 40, 216, 217, 218
  See also Keating et al. (2004), 216
Fowler, C. A., 213, 225, 239
Fox Tree, Jean E., 46, 292
Frank, E., 369
Frank, J. S. *See* Schmidt et al. (1979), 242
Frazier, L., 219
Freedland, M. B. *See* Burnett et al. (1998), 189
Frota, Sónia, 30, 32, 69, 152, 169, 219
Fuchs, Susanne, 32, 227
Fujimura, O., 215
Fujisaki, Hiroya, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 16, 19, 25, 26, 98, 125, 127, 135, 389, 391, 418
  See also Bell et al. (1961), 11, 12, 13
Fujisaki model of prosody, 2–3, 8–11, 13, 19n3, 26–27
  and intonation, 26–27
  superposition in, 5–6
  synthesis of Japanese, 10–11
  and transcription, 19
  and typology, 15
Fujitani, D. S., 184, 188
Fundamental frequency (F0) contours, in models of prosody, 3, 4–5, 52n1
  and intonation, 25
  and typology, 15

Gafos, A., 215, 227n1
  See also Brunner et al. (2014), 215
Gallistel, C. R., 242
Galván, A. *See* Rochet-Capellan et al. (2008), 227
Gandour, J., 384
Gao, M., 216, 221, 222, 223
Garde, P., 276

Gårding, E., 5, 98, 126–127
Garding, Gina, 30, 40, 42, 418
Garellek, M., 165
Garofolo, J. *See* Fiscus et al. (1998), 165
Garrod, S. *See* Anderson et al. (1991), 165
Gaudinat, A. *See* Mertens et al. (2001), 261, 277, 282
Gauthier, B., 401
Gavornik, J. P. *See* Shouval et al. (2014), 242
Gay, T. J., 384
Gee, J. P., 154
Geng, C. *See* Brunner et al. (2014), 215
Gentner, D. R. J., 241
Gerfen, C., 384
German language, 33, 52n3, 134, 157, 218, 248, 249, 251, 290, 291, 292, 307, 308, 358, 364, 365, 367, 368, 413
  AM Tones and Break Indices (G-ToBI) system, 3, 151, 308, 354
  bilingual speakers of, 164
  and bitonal accents, 31
  emphasis in, 305
  Kiel intonation model developed for Northern Standard, 287, 288
  levels of sentence-final pitch, 168
  nuclear accent in, 412
  postlexical pitch accents in, 156
  pronuclear accents in, 48
  rise-fall contours in, 32
  secondary association of phrase accents in, 32
  tonal alignment in, 222
  tonal crowding in, 44
  upstep in, 35
  voiceless coda in, 71
German, James, 39, 248, 412, 413
Germanic languages
  North Germanic languages, 85
  West Germanic languages, 68, 69, 301, 353
Gestures, 213–216, 227
  coordination of, 223–225
  π- gesture model, 216–220
  polysyllabic shortening and rhythm, 225–226
  tone gestures, 221–223
Gibbon, Dafydd, 250
Gibbon, J., 242
Gibson, E., 195n3
  See also Breen et al. (2018), 194; Breen et al. (2012), 171, 189, 194, 196n11; Dilley et al. (2006), 171, 275
Gigi, E. F. *See* Spaai et al. (1993), 262
Gili Fivela, Barbara, 34, 183, 187
  See also Niebuhr et al. (2011), 299, 312
Gilley, Leoma, 79, 242
Glasberg, B. R.52n1
Glewwe, E., 15
Glissando, 263–264
Godfrey, J., 365, 368
Godjevac, Svetlana, 42, 156, 172n6
Goedemans, R., 364
Goldinger, S. D., 364–365
Goldman, J.-Ph., 272
  See also Avanzi et al. (2010), 278fig7.6; Mertens et al. (2001), 261, 277, 282
Goldsmith, John A., 29, 65, 78, 120, 127, 152, 153, 172n4, 182, 183, 184, 186, 187, 188, 190, 191, 195n2, 367
Goldstein, Louis M., 29, 213, 214, 215, 224, 225, 226, 227, 228n4, 228n9, 237
  See also Katsika et al. (2014), 223, 224, 225; Mücke et al. (2012), 216, 222; Nam et al. (2004), 215, 221; Nam et al. (2009), 215; Nam et al. (2008), 225, 226; Niemann et al. (2011), 222; Saltzman et al. (2008), 213, 220, 225, 226, 228, 237, 238
Gooden, Shelome, 31, 32
Goodman, D., 392
Gordon, Matthew, 14, 37, 40, 157, 159
Görs, K., 306
Grabe, Esther, 35, 38, 40, 44, 46, 71, 251, 342–343, 344, 357, 420, 427, 429
Graham, Calbert R., 41
Granlund, C. *See* Nakai et al. (2012), 242
Granström, B., 288
Greek language, 26, 27, 30, 31, 32, 33, 34, 35, 37, 38, 39, 40, 41, 42, 42, 44, 46, 47, 48, 49, 51, 69, 78, 79, 80, 81, 151, 156, 157, 159, 160, 161, 164, 173n17, 182, 194, 208, 209, 210n1, 218, 223, 224, 248, 261
Greenberg, S., 192, 409
Grice, Martine, 7, 30, 31, 32, 33–34, 35, 40, 44, 46, 65, 66, 67, 69, 70, 71, 76–81, 151, 156, 165, 167, 169, 187, 248, 251, 266, 298, 253, 359, 361, 367, 368, 416
  See also Bruggeman et al. (2018), 72; Cangemi et al. (2016), 72; Kügler et al. (2015), 295, 308; Niemann et al. (2011), 222
Grønnum, Nina, 5, 85, 92, 98, 108, 110, 111
Grosjean, F., 154

Groundwork of English Intonation, The (Kingdon), 329
GRToBI representations, 51
  melodies in AM analysis, 26–27
  notations for scaling in, 35
  scaling in, 42, 43
  secondary association of phrase accents in, 32, 33–34
  tonal crowding in, 44
Grudin, J., 241
Gryllia, Stella, 26, 27fig1.1, 35, 47
Gu, Wentao. *See* Fujisaki et al. (2005), 4; Prom-On et al. (2016), 34
Guardiano, C., 400
Gubian, Michele, 81n1
Guenther, F. H., 189
  See also Patel et al. (2011), 189
Guerti, M., 130
Guion-Anderson, S., 399
Gun-wok language, 151, 156, 162, 173n17
Gussenhoven, Carlos, 25, 31, 32, 33, 36, 37, 38, 41, 42, 46, 48, 77, 79, 174n21, 183, 189, 251, 272, 273, 292, 338, 340, 344, 345, 362, 367, 412, 429
Gut, U., 164, 368

Haken, H., 214
Hain, T. C. *See* Burnett et al. (1998), 189
Halle, Morris, 6, 11, 12, 15, 66, 119, 184, 190, 191, 323, 409
Hallé, P. A., 196n17
Halliday, M. A. K., 117, 121, 124, 153, 253, 290, 334, 335, 336, 341
Hancock, P. A., 242
Handbook of Phonetics (Sweet), 323
Hannon, E. E., 188
Hanson, H., 216
Hanssen, J., 429
Hardcastle, W. J., 216
Harrington, J., 165
  See also Miller et al. (1992), 165
Hart, John, 322
Harwood, D. L., 196n17
Hasegawa-Johnson, M., 163, 369
  See also Yoon et al. (2004), 368, 369
Haspelmath, Martin, 50
Hat pattern, 302
Hawkins, B. *See* Schmidt et al. (1979), 242
Hayes, B., 172n3, 191, 192, 415
Hazeltine, R. E., 242
Heffner, 183, 187, 192
Heger, S., 89
Heinz, M. *See* Bell et al. (1961), 11, 12, 13
Heldner, M., 272
Hellmuth, Sam, 31, 32, 39, 69
Herman, Rebecca, 42
Hermes, A., 215
  See also Mücke et al. (2012), 216, 222
Hermes, D. J., 52n1, 262
Hickok, G., 189
Hidden Markov modeling, 36
Hiller, S. M., 238
Hindi language, 65
Hirschberg, Julia, 8, 26, 27fig1.1, 28, 37, 39, 40, 46–48, 51, 68, 71, 151, 154, 159, 166, 172n4, 182, 205, 223, 247, 249, 251, 252, 253, 290, 291, 353, 385, 395, 401, 409, 414, 415, 427
  See also Beckman et al. (2005), 154; Silverman et al. (1992), 50, 117, 136, 137, 151, 247, 338, 414, 418, 427; Syrdal et al. (2001), 368
Hirose, H., 221
Hirose, Keikichi, 2, 3, 9, 10, 98, 418
Hirst, Daniel, 12–13, 51, 117, 118, 119, 120, 123, 124, 125, 130, 134, 135, 136, 137, 139, 141, 143, 145, 163, 196n13, 207, 262, 263, 272, 275, 379
  See also Rossi et al. (1981), 263, 275
Hitchcock, L., 192
Hjelmslev, L., 86
Hockett, Charles F., 8, 427, 428
Hoekstra, J., 183, 196n12
Holliday, Nicole, 71
Holliman, E., 365
Holm, B., 385, 389
Holmes, Frederika. *See* Ashby et al. (1995), 249
Hombert, J.-M., 221, 384
Honda, Kiyoshi, 42
Honorof, D. N., 183, 209, 274
Hoole, P., 216, 218, 230
  See also Bombien et al. (2010), 216, 218
Hoskins, Steven, 38
House, D., 183, 189, 263, 264, 267, 291, 311, 364
House, Jill, 429
Høysgaard, J. P., 85
Hsieh, F.-Y., 222, 223
Hsu, C., 388
  See also Keating et al. (2004), 216
Hualde, José Ignacio, 29, 43, 50, 52, 77, 167, 168, 169, 170, 183, 185
Huang, T. *See* Peng et al. (2005), 156, 159, 166
Huggins, A. W. F., 225

Hume, E. *See* Pitt et al. (2005), 165
Human Communication Research Center (HCRC), 165
Hungarian language, 33, 34, 46
Hunnicut, S., 288
Hussain Shuler, M. G. *See* Shouval et al. (2014), 242
Hutchins, S., 189
  See also Pfordresher et al. (2015), 189, 196n17
Hwang, H. K., 429
Hyman, Larry M., 14, 49, 50, 53n5, 187, 191, 384

Ide, N., 134
Idsardi, W., 190, 191
Igarashi, Y. *See* Maekawa et al. (2002), 182
Iivonen, A., 127
Indjieva, E., 152
Inkelas, Sharon, 20n6, 191
Instituut voor Perceptie Onderzoek (IPO), 263
Intermediate phase (ip), 152
International Phonetic Alphabet (IPA), and prosodic models, 16–19
International Prosodic Alphabet, 18
International Transcription System for Intonation (ITSINT), 51, 117, 141
  mapping from Momel to, 139–140
Interpolation, and phonetic implementation in AM model, 43–44
Interval tier, 155
Intonation, phonological and non-phonological models of, 3
  as an aspect of AM model, 25–26
  defined, 25
  and meaning, 8, 46–48
Intonational meaning, 46–48
Intonational phase (IP), 152
Intonational typology, 48–50
Intonation Curves (Jones), 325
Intrinsic micro-prosody, 265
IPA, and prosodic models, 16–19
Ipek, C., 429
Isard, S. *See* Anderson et al. (1991), 165
IsiZulu language, 139
Italian language, 34, 39, 40, 43, 48, 71, 134, 138, 165, 169, 222
Ironic prosody, 46
Ito, J., 426
Ito, K., 165
Iversen, J. R., 273
Ivry, R., 242

Jacobson, R. *See* Glewwe et al. (2018), 15
Jaeger, T. Florian. *See* Beaver et al. (2007), 39, 46, 412, 416
Jakobson, Roman, 20n7, 184
Jamaican Creole language, 31, 32
Jannedy, S. *See* Kügler et al. (2015), 295, 308
Japanese language, 3, 9, 12, 33, 48, 49, 127, 288, 385, 408, 415, 416, 418, 419, 420
  accent and phrase commands in, 4
  final lowering in, 42
  and the Fujisaki model, 10–11
  fundamental frequency contours, in models of prosody, 5–6
  intonational typology and, 49
  metrical structure of, 37
  and secondary association of phrase accents, 33
  synthesis of, 10
  ToBI representation for, 51, 52n2, , 151, 153, 155, 156, 158, 159, 160, 161, 163, 164, 165, 173n12, 248
  underspecification in, 44
Jarvis, E., 396
Jarzabkowska, P. *See* Niebuhr et al. (2012), 292, 308
Jaskula, Marek, 26, 35, 39, 44–45
Jassem, W., 117, 121–123, 124, 272
Jensen, C., 85, 110
Jerndorff, P., 85, 86, 91–92, 95
Jespersen, Otto, 86
Jesus, M. T., 318
Jilka, M., 357
Johnson, K., 357
  See also Pitt et al. (2005), 165
Jones, Daniel, 290, 305, 325
Jones, L. A., 242
Jones, M. R., 188, 195n6, 196n17
Jun, Sun-Ah, 25, 30, 31, 32, 37, 40, 48–49, 50, 51, 64, 65, 71, 77, 110, 117, 151, 156, 159–160, 162, 163, 164–166, 169, 172n1, 172n2, 172n3, 173n12, 183, 184, 185, 194, 213, 216, 218, 225, 367, 420
  See also Delais et al. (2015), 152
Jurafsky, D. *See* Calhoun et al. (2010), 353, 365

Kainada, Evia, 40
  See also Baltazani et al. (2015), 47, 48
Kalbertodt, Janina, 69
Kamen, R. S., 396
Kaminskaia, S., 272
Kanté, M., 124

Karlsson, Anastasia, 28, 37, 163
Katsika, Argyro, 37, 77, 79, 81, 218, 223, 224–225, 227
Katsumata, H., 241
Katz, J., 193, 429
Kaufmann, Stefan, 39, 412, 413
Kaun, Abigail. *See* Byrd et al. (2000), 38, 213, 216
Kawahara, Shigeto, 50
Keane, Elinor, 37
Keating, Patricia A., 38, 183, 209, 216, 217, 218, 230, 418
Kelly, A., 388, 401
Kelly, N. E. 265
Kelso, J. A. S., 214, 221, 382
Kennedy, C., 413
Kenworthy, J., 166, 329
Kerkhoff, Joop. *See* Gussenhoven et al. (2003), 251
Kerzel, D. *See* Mechsner et al. (2001), 382
Khan, Sameer ud Dowla, 37, 51, 152, 156
Kiel Corpus of Spontaneous Speech, 298
Kiel Intonation Model (KIM), 287–288, 293, 313
  concatenation contour, 302–303
  emphatic intonation, 304–308
  further developments in, 309–313
  history of, 288–290
  methodological guideline, 290–292
  perception experiments, 292
  phase-final contour, 303–304
  phase-initial contour, 297
  phonological elements, 293
  pitch-accent contour, 297–301
  prominence levels, 293–295
  tune structure and phrase boundaries, 295–297
  working with, 308–309
Kiesling, S. *See* Pitt et al. (2005), 165
Kikuchi, H. *See* Maekawa et al. (2002), 182
Kilbourn-Ceron, Oriana, 71
Kim, H. *See* Hasegawa-Johnson et al. (2005), 369
Kim, Keeho. *See* Jun et al. (2000), 71
Kim, S.-S. *See* Hasegawa-Johnson et al. (2005), 369
Kim, B., 166, 170
Kinetic tones, 329–330
King, S. *See* Taylor et al. (1999), 357
Kingdon, Roger, 117, 142, 253, 329–331
Kipare language, final lowering in, 42
Kiparsky, Paul, 14, 249
Kirsch, A., 309
Kjelgaard, M. M., 165
Klatt, D. H., 154, 225
Kleber, F., 288
Klouda, G. V. *See* Eady et al. (1986), 399
Knight, Rachel-Anne, 44, 183, 189
Knoblich, G. *See* Mechsner et al. (2001), 382
Koch, G. G., 370n7
Kochanski, G., 418, 427
Koenig, Laura L. *See* Fuchs et al. (2015), 32
Kohler, Klaus J. 287–290, 291–292, 294, 295–296, 297, 298, 300, 301, 302, 303–304, 305–306, 308, 310, 353, 367
Kohonen, T., 401
Koike, C., 165
Kollmeier, B., 356
Korean language, 65, 77, 135, 136, 139, 151, 155, 158, 160, 161, 163, 164, 166, 172n2, 217, 218
  intonational encoding of focus in, 48–49
  K_ToBI representations, 51
  metrical structure of, 37
  multitonal combinations of edge tones in, 31
  and pitch accents, 37
Kraemer, J. *See* Breen et al. (2012), 171, 189, 194, 196n11; Dilley et al. (2006), 171, 275
Krahmer, Emiel, 39, 69, 227
Krakow, Rena Arens, 215, 417
Kreiman, J., 239
Krifka, M., 193
Kristoffersen, G.
Krivokapić, J., 216, 217, 218, 219, 220, 224, 225, 237, 238, 24
  See also Katsika et al. (2014), 223, 224, 225; Nam et al. (2008), 225, 226; Saltzman et al. (2008), 213, 220, 225, 226, 228, 237, 238
Kruckenberg, A., 239, 293
Krüger, Martina, 71
Krumhansl, C. L., 197
Kudela-Dobrogowska, K., 272
Kügler, Frank, 46, 69, 295, 308, 353, 429
Kuhl, P. K., 396, 401
Kühnert, B. *See* Bombien et al. (2010), 216, 218
Kunnari, Sari. *See* Nakai et al. (2009), 38, 242; Nakai et al. (2012), 242
Kuo, G., 170

Laboissière, R. *See* Rochet-Capellan et al. (2008), 227
Lacerda, F., 365, 266
  See also Kuhl et al. (1992), 401
Lacheret-Dujour, A.
Ladd, D. Robert, 3, 4, 8, 15, 19n4, 25, 26, 27, 28, 29, 30, 31, 32, 33–34, 35, 36–37, 38, 39, 40, 41, 42, 43, 44, 48, 50, 52, 65, 71, 76, 77, 78, 81, 87, 104, 120, 151, 152, 154, 157, 164, 166, 167, 182, 183, 187, 191, 192, 194, 196n14, 208, 213, 219, 222–223, 247, 248, 250, 251, 252, 266, 272, 274, 276, 291, 291, 295, 320–321, 336, 338, 339, 346n7, 352, 367, 368, 384, 395, 412, 416, 430
  See also Grice et al. (2000), 7
Ladefoged, Peter, 15, 20n9
Ladusaw, William A., 20n10
Lahiri, A., 415
Lambrecht, K., 111
Landgraf, R., 287, 292, 306
Laniran, Yetunde O., 42, 384
Langston, A. *See* Brugos et al. (2019), 19
Large, E. W., 195n6
Larson, C. R. *See* Burnett et al. (1998), 189; Chen et al. (2007), 89
Latash, M. L., 382
Leading tone, 153
Lebanese Arabic language, 32
Leben, William R., 65, 158, 172n4, 191
Lee, Albert. *See also* Xu et al. (2015), 25, 29, 40, 44, 402; Xu et al. (2013), 401
Lee, C.-Y.183, 184
Lee, E. K., 164, 165, 170
Lee, O. J. *See* Peng et al. (2005), 156, 159, 166
Lee, K. S. *See* Syrdal et al. (2000), 368
Lee, Sook-hyang. *See* Jun et al. (2000), 71
Lee, Yong-Ju. *See* Jun et al. (2000), 71
Lefkowitz, L. M., 242
Lehiste, I., 154, 225
Lengeris, Aggelos. *See* Baltazani et al. (2015), 47, 48
Leonard, T., 241
Liberman, Mark Y., 19n4, 42, 65, 107, 120, 152, 154, 172n4, 182, 184, 185, 388, 408, 416, 418
Lickley, Robin, 34, 40, 41, 164
Lindblom, B. *See* Kuhl et al. (1992), 401
Lindqvist, J., 12
Lintfert, B., 369
Litman, Diane, 414, 415
Liu, Fang, 8
  See also Xu et al. (2015), 25, 29, 40, 44, 402
Liu, H. *See* Chen et al. (2007), 89
Liu, X. *See* Xu et al. (2013), 401
Loakes, Debbie. *See* Fletcher et al. (2016), 77
Local, John, 79
Locke, J. L., 396
Loevenbruck, Hélène, 40, 292
  See also D'Imperio et al. (2007), 223
Lohfink, Georg, 77, 79, 80fig1r.2, 81
Longobardi, G., 400
Lorenz, U. *See* Niebuhr et al. (2012), 292, 308
Lotts, D. W. *See* Eady et al. (1986), 399
Loucks, T. M., 189
Loui, P. *See* Pfordresher et al. (2015), 189, 196n17
Lund, J., 85
Lupyan, G. *See* Dingemanse et al. (2015), 415
Lützen, Peter H., 85

MacNeilage, P. F., 415
Macromelody, 127–130
  profiles, 130
Macro-rhythm, 49
McAuley, J. D., 194
  See also Breen et al. (2014), 194; Miller et al. (1992), 165
McCarthy, John, 19n3, 195n6
McClean, M., 216
McDaniel, J., 365, 368
McGory, J. T., 164, 368, 369
  See also Beckman et al. (2002), 152; Syrdal et al. (2001), 368
McGowan, R. S., 221
McQueen, J., 217, 218
Maday, K., 165
Maeda, Shinji, 42
Maekawa, Kikuo, 172n7
Maghbouleh, A., 135–136
Mahrt, T., 170
Makashay, M. J. *See* Syrdal et al. (2000), 368
Maltese language, 65, 66, 67, 68, 76, 77
Mandarin Chinese language, 4, 124, 399
Maneewongvatana, S. *See* Chuenwattanapranithi et al. (2008), 388, 401
Manfredi, V., 192
Map Task (HCRC), 165
Marin, S., 215
Marjorie, C. *See* Peng et al. (2005), 156, 159, 166

Martin, P. *See* Rossi et al. (1981), 263, 275
Martinet, André, 8
Maskikit-Essed, Raechel, 77
Matthies, M. L., 241
Mattys, S., 194
Mayer, Jörg, 354, 357, 358, 359, 361, 366–367
See also Grice et al. (1996), 248, 251, 368
Mayo, Catherine, 71, 151, 251, 368
Mayo, N. *See* Calhoun et al. (2010), 353, 365
Mechsner, F., 382
Meir, I., 415
See also Sandler et al. (2011), 8
Meltzoff, A. N., 196
Menezes, C. *See* D'Imperio et al. (2007), 223
Menn, Lise, 42
Mennen, Ineke, 29, 30, 33, 34, 40, 41, 44, 77, 78, 157, 164, 182, 192, 194
Merchant, H. W., 242
Mertens, Jane, 69
Mertens, P., 189, 261, 262, 263–264, 267, 270, 272, 273, 274, 275, 276, 277
Mester, A., 416
Metrical structure, 37–39
Micromelodic effects, 125–127
Micromelody, 127–130
Microprosody, 87
Micro-rhythm, 49
Mikkelsen, K., 86
Miller, J., 165
Minnett, J. *See* Peng et al. (2012), 183
Minus accent, 301
Miscellaneous tier, 161
Mitterer, H. *See* Dingemanse et al. (2016), 415
Mixdorff, Hans-Jörg, 3, 10, 19, 20n5, 135, 310
Mo, Y., 163, 170
Möbius, B., 98, 293, 353, 356, 368, 369
See also Barbisch et al. (2007), 358; Schweitzer et al. (2015), 366; Wade et al. (2010), 366
Model-based parametric representation, 430
Möhler, G., 351, 357, 358, 369
Momel, 133–136
Monelle, R., 184
Moore, Brian, 52n1
Mooshammer, C., 218, 220
See also Bombien et al. (2010), 216, 218; Katsika et al. (2014), 223, 224, 225
Monaghan, P. *See* Dingemanse et al. (2015), 415
Mongolian language, 29, 37, 152, 158, 163, 172n2
Moreton, Elliott, 14–15
Morgan, T. A. *See* Beckman et al. (2002), 152
Morrill, T., 194
See also Miller et al. (1992), 165
Morton, R., 336
Moulton, William G., 249
Mücke, D., 216, 222, 228n8
See also Niemann et al. (2011), 222
Mueller, P. R., 399, 429
See also Eady et al. (1986), 399
Mulac, A., 192
Müller, F., 308
Munhall, K. G., 213, 221, 237
Mushin, I. *See* Stirling et al. (2001), 165
Myers, Scott, 40, 195n9, 196–197n18

Nagashima, S., 9
Nakai, Satsuki, 38, 79, 80fig1r.2, 242
Nakatani, L. H., 225
Nam, Hosung, 213, 214, 215, 221, 226, 227, 237
See also Goldstein et al. (2009), 224; Mücke et al. (2012), 216, 222; Niemann et al. (2011), 222; Prom-On et al. (2016), 34; Saltzman et al. (2008), 213, 220, 225, 226, 228, 237, 238
Narayanan, Shri, 369
See also Byrd et al. (2000), 38, 213, 216; Byrd et al. (2009), 215
Nash, R., 292
Nature and natural in phonetics, 14–15
Navarrete, E. *See* Dupoux et al. (2008), 164
Nespor, Marina, 38, 120, 152, 154, 172n3, 173n13
Newell, K. M., 242
Newman, D., 305
Nguyen, N. *See* D'Imperio et al. (2007), 223
Nibert, Holy, 42
Ní Chasaide, Ailbhe, 40
Nichols, J., 400
Nicolas, P., 263
Niebuhr, Oliver, 25, 39, 48, 69, 183, 189, 196n12, 291, 292, 295, 297, 298, 299, 300, 301, 302, 306, 307, 308, 309, 310, 311, 312
See also Kügler et al. (2015), 295, 308
Niemann, Henrik, 222
See also Grice et al. (2017), 71
Nieminen, T., 226, 238
Nikolaidis, Katerina. *See* Baltazani et al. (2015), 47, 48

Ning, L. H., 189
Nishinuma, Y. *See* Rossi et al. (1981), 263, 275
Niziolek, C. *See* Patel et al. (2011), 189
Noble, L., 388, 401
Nolan, Francis, 44, 51, 52n1, 71, 76, 183, 262, 342, 343, 346n7
  See also Grabe et al. (2000), 40, 44, 429
Nooteboom, S., 125–126
Nord, L., 239
Nucleus, in AM, 36, 48

O'Brien, M., 164
O'Connor, J. D., 46, 117, 142, 290, 324, 331–333, 334, 335, 336, 340
O'Connor, K., 225
O'Connor, R. J., 184
Odden, D., 184, 190
O'Dell, M. L., 226, 238
Oh, M., 164
Ohala, John, 46, 221
Öhman, S. E. G., 12, 98
Ohno, S., 5, 16
  See also Fujisaki et al. (2005), 4
Ohta, K., 98
Oliveira, Pedro, 177
  See also Frota et al. (2015), 152
Oller, K. D., 216
Omlæggertone, 85
Optimality theory, 14
Orlikoff, R. F., 262
Ortega-Liebaria, M., 169
Osgood, C. E., 292
Ostendorf, Mari, 36, 38, 165, 353, 412
  See also Price et al. (1991), 154, 247; Silverman et al. (1992), 50, 117, 136, 137, 151, 247, 338, 414, 418, 427; Wightman et al. (1992), 54, 247
Oxenvad, E., 85, 86

Paccia-Cooper, P., 154, 220
Padden, C. *See* Meir et al. (2013), 415; Sandler et al. (2011), 8
Padget, J., 416
PaIntE model of intonation, 351–352, 376–370
  and intonation research, 363–367
  parameters, 352–358
  and prosodic categories, 358–363
Pallet, D. *See* Fiscus et al. (1998), 165
Palmer, Harold E., 117, 153, 253, 326–329, 330, 331, 334, 337, 346n5
Patel, A. D,. 184, 188, 272
Patel, R., 189
Pater, Joe, 15
Patil, U., 429
Pearson, D. M., 396
Peng, G., 183
Peng, S. H., 156, 159, 166, 398
PENTA model, 377, 408–410, 420–421, 425–428, 431
  broader significance, 391–394
  and cognition, 396–397
  and computational modeling, 394–396
  conceptual framework, 377–388
  controlled experiments, 428–430
  economy of representation, 430–431
  encoding schemes as prosophemes in, 400–401
  generalization, 418–420
  and intonation, 26–27
  and perception, 401
  and phonology, 397–399
  phonology and phonetic realization, 415–418
  and prosodic typology, 399–400
  as a research tool, 388–391
  and semantics versus phonology, 410–415
  and tonal crowding, 45
Peperkamp, S. *See* Dupoux et al. (2008), 164
Peretz, I., 189
Perkell, J. S., 241
Perlman, M., 197n19
Peters, Benno, 287, 288, 301, 303, 308
Peters, Jörg, 48, 367, 429
Petersen, N. R., 102
Petkov, C. I., 396
Petrone, Caterina, 40, 48, 69, 291
  See also Fuchs et al. (2015), 32
Pfitzinger., P., 310
Pfordresher, P. Q., 189, 196n17
Phonetic categories, and transcription, 16–19
Phonetic implementation, and prosodic models, 9–10
  in AM model, 39–45
  and phonetic variability, 77–79
Phonetic representation, 118–119, 125–136
Phonological model of prosody, 3–7
  and diphones, 13–14
Phonology
  and prosodic theory, 1
  and psychological realtiy, 15–16
Piccinini, A., 79

Pierrehumbert, Janet, 7, 8, 25, 28, 29–30, 31, 32–33, 34–36, 37, 38, 39, 40, 41–42, 43, 44, 46–48, 51–52, 65, 68, 76, 78, 107, 120, 143, 152, 158, 159, 166, 172n4, 182, 184, 185, 186, 187, 188, 189, 192, 193, 194, 195n9, 196n10, 196n14, 204–205, 206, 207, 219, 247, 250–251, 331, 338, 350, 351, 353, 365–366, 379, 385, 395, 396, 397, 401, 408, 409, 411, 412, 413, 414, 416, 417, 418, 419, 420, 425, 427, 428, 429, 430
  See also Silverman et al. (1992), 50, 117, 136, 137, 151, 247, 338, 414, 418, 427
Pike, Kenneth, 34, 52–53n4, 142, 143, 225, 249, 274
Pitch course, 87
Pitch level and pitch span, in AM model, 41
Pitch range, 271–272
Piterman, Michel, 43–44
Pitrelli, John, 71, 247, 252, 368, 419
  See also Silverman et al. (1992), 50, 117, 136, 137, 151, 247, 338, 414, 418, 427
Pitt, M. A., 165
  See also Miller et al. (1992), 165
Poiré, F., 272
Polish language, 26–27, 35, 39, 44, 45
Polysyllabic shortening, 225–226
Popper, A., 188, 196n17
Port, R., 214, 225, 226
Portuguese language, 30–31, 32, 63, 152
Post, Brechtje, 51, 71, 342, 420
  See also Delais et al. (2015), 152; Grabe et al. (2000), 40, 44, 429
Postfocus compression (PFC), 399
Postlexical prosodic events, 48–49
Pouplier, M., 215, 227–228n1
  See also Goldstein et al. (2009), 224
Prado, L., 242
  See also Merchant et al. (2008), 242
Price, Patti, 36, 154, 165, 247, 353
  See also Silverman et al. (1992), 50, 117, 136, 137, 151, 247, 338, 414, 418, 427; Wightman et al. (1992), 54, 247
Prieto, Pilar, 26, 31, 34, 39, 40, 42, 43, 44, 47, 50, 52, 60, 77, 78fig1r.1, 79, 152, 168, 169, 169, 183, 185, 187, 205, 223, 227, 248
  See also D'Imperio et al. (2005), 37; Escudero et al. (2012), 71
Primer of Phonetics (Sweet), 323
Prince, Alan, 14, 15, 38, 120
Prinz, W. *See* Mechsner et al. (2001), 382
Proctor, M., 216
PROLAB, 288
Prom-on, Santitham, 8, 44, 382, 384, 385, 388, 389, 390, 391, 394, 398, 399, 401, 408, 409, 410, 411, 413, 414, 415, 416, 417, 419, 420, 428, 430, 431
  See also Liu et al. (2013), 22; Liu et al. (2013), 8, 382, 387, 390, 391, 399, 400, 428, 429, 430; Xu et al. (2015), 25, 29, 40, 44, 402
Prosodic licensing, 415
Prosodic Systems and Intonation in English (Crystal), 336
Prosogram model for pitch stylization, 259–262, 280–282
  intonation, automatic symbolic transcription of, 276–280
  pitch level and movement, transcription of, 273–276
  prosodic features obtained during stylization, 271–273
  stylization algorithm, 266–271
  and a tonal perception model, 262–263
  and tonal perception, 263–266
Prosophemes, 400, 401
ProZed, 141
Pruitt, K., 414
Przybocki, M. *See* Fiscus et al. (1998), 165
PSOLA resynthesis, 133–134
Psychology, and prosodic theory, 2, 15–16
Pullum, Geoffrey K., 20n10

Quarmby, Louise May. *See* Ladd et al. (2009), 40, 78
Quinn, J. T. *See* Schmidt et al. (1979), 242
Quirk, R., 336

Rahmani, H., 401
Rakerd, B., 225, 249
Ramanarayanan, V.
Rapid prosody transcription system (RPT), 170
Rapp, S., 188, 353, 396
RaP prosodic transcription system, 188–195, 207–210
Rathcke, Tamara, 44, 69, 71, 287
Rauzy, S., 136
Raymond, W. *See* Pitt et al. (2005), 165
Reciprocal property, 190
Redi, Laura, 26
Reed, Paul, 71
Refice, Mario, 70–71

Reichel, Uwe D. *See* Fuchs et al. (2015), 32
Reilly, K. *See* Patel et al. (2011), 189
Reinisch, E. *See* Dingemanse et al. (2016), 415
Remijsen, Bert, 40, 79, 242
Reverse tone (Danish), 85
Reyelt, Matthias. *See* Grice et al. (1996), 248, 251, 368
Rhythm units, 121–124
Rialland, Annie, 46, 400, 415, 420
Rice, Keren D., 192
Ridouane, Rachid, 40, 46, 65
Riester, A., 353
Rietveld, Toni, 42, 46, 271, 272, 292, 362
  See also Gussenhoven et al. (2003), 251
Riggs, D., 223, 224, 225
Rischel, J., 89
Rise/fall/connection (RFC) model of intonation, 134
Ritchart, Amanda, 46
Ritter, Simon, 32, 310
  See also Grice et al. (2017), 71
Rochet-Capellan, Amélie, 227
  See also Fuchs et al. (2015), 32
Roelofsen, F., 414
Roettger, Timo B., 40, 46, 65, 66, 67, 71
  See also Grice et al. (2017), 71
Röhr, C., 353
Romanian language, 33, 46
Romanzi tier, 156
Rooth, M., 411
Roseano, P., 169
Rosenbaum, D. A., 252, 417
Rosenberg, A., 368
Rosenberg, J. C., 273
Ross, Ken, 38, 412
Rossi, M., 263, 265, 266, 275
Rump, H. H., 429
Rusaw, E., 223, 224, 228n10
Russell. D. M., 241
Russian language, 11, 12, 44, 71, , 124, 139, 174n21, 249, 409
Rutkowski, J. *See* Pfordresher et al. (2015), 189, 196n17

Sag, Ivan, 19n4
Salihie, M. *See* del Giudice et al. (2007), 183
Saltzman, Elliott, 38, 213, 214, 215, 216, 217, 218, 220, 221, 225, 226, 227, 238, 242
  See also Byrd et al. (2009), 215; Goldstein et al. (2009), 224; Goldstein et al. (2009), 224; Nam et al. (2004), 215, 221; Nam et al. (2009), 215; Nam et al. (2008), 225, 226
Sanderman, A. A., 219
Sanders, L. D. *See* Breen et al. (2014), 194; Breen et al. (2010), 193
Sandler, W., 8
  See also Meir et al. (2013), 415
Sapir, E., 15
Saussure, 20n7, 289, 385, 425, 427, 428
Savino, Michelina, 70, 71, 165
  See also Grice et al. (2005), 165, 168, 169
Säuberlich, B. *See* Barbisch et al. (2007), 358
Scaling, and AM model, 29, 34–35
  and phonetic implementation in AM model, 41–43
Schafer, A., 165
Scheffers, M. T. M., 262
Schepman, Astrid, 30, 34, 40, 41, 43, 78, 164, 183, 187, 192, 353
  See also Ladd et al. (1999), 78; Ladd et al. (2009), 40, 78
Schmidt, R. A., 252
Schroeder, B., 295
Schroeter, J. *See* Syrdal et al. (2000), 368
Schuerman, W. *See* Dingemanse et al. (2016), 415
Schuh, 187
Schulz, C. *See* Niebuhr et al. (2012), 292, 308
Schütze, H. *See* Schweitzer et al. (2015), 366; Wade et al. (2010), 366
Schwartz, J. L. *See* Rochet-Capellan et al. (2008), 227
Schwarzschild, R., 412, 413
Schweitzer, A., 253, 364–365, 366, 368, 369
  See also Barbisch et al. (2007), 358; Schweitzer et al. (2015), 366
Schweitzer, K., 365, 366–367
Scobbie, James M., 79, 80fig1r.2
Scott, D., 154
Searle, J., 195n3
Sebastián-Gallés, N. *See* Dupoux et al. (2008), 164
Secondary association, 32–34
Segmental anchoring, 40
Selkirk, Elisabeth O., 37, 38, 46, 71, 120, 152, 154, 172n3, 193, 220, 429
Semjen, A. , 241
Sennett, W., 225, 239
Serbo-Croatian language, 151, 157, 163, 172n6
Seyfarth, R. M., 8
Shaffer, L. H., 241

Shafran, I. *See* Ostendorf et al. (2001), 165
Shannon, C. E., 197
Sharpe, V., 194
Shattuck-Hufnagel, Stefanie, 36, 38, 51, 71, 151, 154, 165, 166, 167, 170, 172n4, 174n21, 182, 205, 210, 213, 216, 223, 224, 225, 238, 240, 243, 247, 249, 251, 353, 412
  See also Barnes et al. (2010), 34; Barnes et al. (2012), 43, 72, 182, 183, 189; Barnes et al. (2013), 43; Barnes et al. (2014), 266; Beckman et al. (2005), 154; Brugos et al. (2008), 162; Ostendorf et al. (2001), 165; Price et al. (1991), 154, 247; Wightman et al. (1992), 54, 247
Shaw, Jason A., 50
Shi, R., 401
Shih, Chilin, 20n6, 42, 189
Shilluk language, 40
Shimron, J., 195n6
Shosted, R. *See* del Giudice et al. (2007), 183
Shouval, H. Z. , 242
Shue, Y.-L., 429
Sichel-Bazin, Rafèu, 47
Siegel, J. A., 196n17
Siegel, W., 196n17
Silverman, Kim, 36, 40, 41, 44, 50–51, 117, 136, 137, 151, 223, 247, 338, 384, 414, 417, 418, 427
Šimko, J., 239
Simon, A. C., 272, 367
  See also Avanzi et al. (2010), 278fig7.6; Goldman, Auchlin, et al. (2007), 272
Simpson, Adrian P., 79
Singer, Ruth, 242
  See also Fletcher et al. (2016), 77
Sityaev, D., 429
Slope changes, 193–194
Smalley, W. A., 275
Smiljanic, Rajka, 40, 365
Smillie, C., 388, 401
Smith, Henry Lee, 34, 206, 408
Smolensky, Paul, 14, 15
Smolibocki, B. *See* Kügler et al. (2015), 295, 308
Snedeker, J., 165, 219
Sneedorff-Birch, F., 85
Snider, K., 191
Sodigov, F. *See* Niebuhr et al. (2012), 292, 308
Sonderegger, Morgan, 71
Sorensen, J. M., 362
Sotiropoulou, S. *See* Brunner et al. (2014), 215
Southard, D. L., 382
Spaai, G. W. G., 262
Spanish language, 37, 39, 40, 42, 47, 49, 69, 70, 71, 77, 134, 139, 157, 168, 169
Spectral stability hypothesis, 263
Speer, S. R., 165, 166
  See also Schafer et al. (2000), 165
Spencer, R. M., 241, 242
Spini, M. *See* Loevenbruck et al. (2013), 292
Sproat, R., 215, 253
Sridhar, V. K. R., 369
Stackhouse, Rebekah. *See* Ladd et al. (2009), 40, 78
Steedman, Mark, 47, 253
  See also Calhoun et al. (2010), 353, 365
Steele, J., 322–323, 324fig9.1, 325, 339, 346n3
Steele, S. A., 189
Stelmach. G. E., 241
Stevens, Kenneth N., 11, 12, 216, 385
  See also Bell et al. (1961), 11, 12, 13; Kuhl et al. (1992), 401
Stirling, L., 165
Stoakes, Hywel. *See* Fletcher et al. (2016), 77
Storm, A. *See* Spaai et al. (1993), 262
Strangert, E., 308
Streeter, L. A., 416
Stress foot, 121
Strom, V. *See* Syrdal et al. (2000), 368
Strömbergsson, S., 291
Strunk, Klaus, 71
Stylianou, Y. *See* Syrdal et al. (2000), 368
Stylization, 262–263
Suci, G. J., 292
Sun, X., 139, 379, 401, 429
Sundberg, J., 379
Suomi, Kari. *See* Nakai et al. (2009), 38, 242; Nakai et al. (2012), 242
Superposition, in the Fujisaki model, 5–6, 98
Suprasegmentals, 25
Swadesh, M., 427
Swedish language, 32, 85, 88, 89, 90, 92, 94, 98, 111, 158, 238, 291, 285, 417, 420
  See also Danish language, intonation in
Sweet, H., 323–326, 337
Swerts, Marc, 39, 69, 219, 227
Synthesis
  and phonetic implementation, 9–13
  quality of, as criterion for model choice, 13–14
Syrdal, A.K.,378, 379

Tabain, M., 216
Taheri-Ardali, M., 401
Tamil language, metrical structure of, 37
Tannenbaum, P. H., 292
Tashlhiyt Berber language, 40, 46, 65, 66, 67, 68
Task Dynamics application model (TADA), 215–216
Taylor, Paul, 117, 133, 262, 351, 353, 357, 379, 391
TBU, 29
  and bitonal accents, in Greek, 31
  and secondary association of phrase accents, 32–33
Tees, R. C., 401
Terken, Jacques, 43–44, 353
  See also Gussenhoven et al. (2003), 251
Ternes, E., 17
't Hart, Johan, 40, 42, 46, 78, 98, 125, 137, 196n14, 262, 263, 264, 266, 271, 273, 274, 275, 427
Themistocleous, Charalambos, 38
Thipakorn, B., 382, 388, 389, 390, 401, 428, 430
  See also Chuenwattanapranithi et al. (2008), 388, 401
Thorsen, N. G., 85, 88, 110, 167
Tiede, M. *See* Katsika et al. (2014), 223, 224, 225
Tierney, A., 197n19
Tilsen, S., 222, 226
Timing slots, 192–193
Tobin, S. *See* Byrd et al. (2009), 215
ToBI systems, family of, 18, 50–52, 151–152, 171–172
  Alt (alternative) tier, 162
  break indices tier, 160–161
  challenges and recent developments, 166–171
  conventions, 154–155
  and gradiance, 252–254
  miscellaneous tier, 161
  other tiers, 162
  and phonological analysis of intonation, 249–252
  and prosodic typology, 205–206
  strengths of, 162–166
  theoretical background, 152–154
  tones tier, 156–160
  trouble with, 247–253
  uncertainty, ToBI labels for, 162
  words tier, 155–156
Tonal crowding, 44–45
Tonal units, 121–124
Tøndering, J., 92, 110, 111
Tone-bearing unit (TBU), 29
  and bitonal accents, in Greek, 31
  and secondary association of phrase accents, 32–33
Tone gestures, 221–223
Tones, 29
Tones tier, 156–160
Torreira, Francisco, 26, 40, 69, 77, 81n1, 223
Torres-Tamarit, F. J. *See* Prieto et al. (2009), 152
Trager, George L., 34, 206, 408
Trailing tone, 153
Trainor, L. J., 188
Transcription systems, and prosodic theory, 2, 16–19
Trask, R. L., 305
Treisman, M., 242
Trim, J. L. M., 333, 343
Trofimovich, P., 164
Troost, J., 191
Trubetzkoy, N. S., 119, 289
Truckenbrodt, Hubert, 42, 191
Truncation, 44
Tseng, C.-Y. *See* Peng et al. (2005), 156, 159, 166
Tufvesson, S. *See* Dingemanse et al. (2016), 415
Turk, Alice, 38, 213, 216, 223, 224, 225, 238, 240, 429
  See also Nakai et al. (2009), 38, 242; Nakai et al. (2012), 242
Turkish language, 65, 417, 429
Turvey, M. T., 214
Typology, and prosodic theory, 2, 14–15, 76–77
  intonational, 48–50, 79–81
  and phonetic variablity, 77–79

Ueyama, M., 164
Undergeneration, of language patterns, 14
Undershoot, 44
Underspecification, and phonetic implementation in AM model, 44–45
Universal phonetic categories, 16–18
Upstep, 35
Uptalk, 46

Vaissière, J., 98
Vaknin, V., 195n6
Vallduvi, E., 412
van Donselaar, W., 184

van Gelder, T., 214
van Gestel, J. C., 52n1, 262
Van Lancker, D., 239
van Lieshout, P.
van Ouden, D.-B., 194
Vanrell, Maria del Mar, 42
See also Escudero et al. (2012), 71
van Rijsbergen, C. J., 145n4
van Santen, Jan, 39, 40, 223, 353, 356
van Wieringen, P. C., 241
Vazquez-Larruscaín, M., 85, 111
Veilleux, Nanette, 51, 210, 247
See also Barnes et al. (2010), 34; Barnes et al. (2012), 43, 72, 182, 183, 189; Barnes et al. (2013), 43; Barnes et al. (2014), 266; Brugos et al. (2008), 162
Vella, Alexandra, 66, 67, 68
Venditti, Jennifer J., 25, 49, 50, 51, 5 2, 53n5, 151, 156, 167, 248
See also Maekawa et al. (2002), 182
Verbmobil project, 288
Vermillion, P., 271, 272, 411
Véronis, J., 134
Vigário, Marina, 57
See also D'Imperio et al. (2005), 37; Frota et al. (2015), 152
Vinke, L., 194
Vladimirova-Buhtz, T., 17
Vogel, Irene, 38, 120, 152, 154, 172n3, 173n13, 220
von Thun, F. S., 305
Vonwiller, J. *See* Miller et al. (1992), 165
Vos, P., 191

Wade, T., 366, 427
Wagner, M., 219, 220, 386
See also Breen et al. (2010), 193
Wagner, P., 239, 293
Wahlster, W., 288, 358
Wakefield, J., 124
Wales, R. *See* Stirling et al. (2001), 165
Wallace, A., 384, 385
Walsh, M., 366
See also Schweitzer et al. (2015), 366; Wade et al. (2010), 366
Wang, B., 399, 400, 401
Wang, C., 5, 16, 388
See also Fujisaki et al. (2005), 4
Wang, M., 386
Wang, Q. E., 130
Wang, W. E. 195n1, 195n8
Wang, W. S.-Y., 398
See also Peng et al. (2012), 183
Ward, Gregory, 26, ***27***, 252, 414
Ward, I. C., 117, 142
Warren, Paul, 38, 40, 41, 46, 52n1, 165, 253, 420
See also Schafer et al. (2000), 165
Watson, B. C., 396
Watson, D., 165
Wearden, J. H., 242
Weenink, D., 108, 133–134, 263
Wehrle, Simon. *See* Bruggeman et al. (2018), 72; Cangemi et al. (2016), 72
Wehrli, É. *See* Mertens et al. (2001), 261, 277, 282
Welby, Pauline, 40, 69, 183
See also D'Imperio et al. (2007), 223
Welch, G. F. *See* Pfordresher et al. (2015), 189, 196n17
Wells, J. C., 117, 250, 335, 336
Werker, J. F., 401
Wesener, T., 301
West Greenlandic language, 37, 51
Whalen, Douglas H., 183, 209, 274
See also Prom-On et al. (2016), 34
White, Laurence, 223, 224, 225, 239, 240, 429
See also Ladd et al. (2009), 40, 78
White, S. *See* Schafer et al. (2000), 165
Wightman, Colin, 118, 136, 154, 216, 247
See also Silverman et al. (1992), 50, 117, 136, 137, 151, 247, 338, 414, 418, 427; Syrdal et al. (2000), 368
William of Ockham, 182
Williams, B., 238
Williams, K. A. *See* Kuhl et al. (1992), 401
Winchester, S., 291
Windmann, A., 239
Wing, A. M., 242
Witten, I. H., 369
Włodarczac, M., 272
Wolters, Maria. *See* Beaver et al. (2007), 39, 46, 412, 416
Wong, P. C. M., 184
Wong, Wai Yi P., 51, 156, 157, 159, 173n13
Wu, X. *See* Xu et al. (2013), 401

Xu, C. X., 195n1, 382, 384, 399, 401, 429, 431
Xu, Yi, 8, 25, 26–27, 40, 44, 46, 139, 195n1, 195n8, 221, 273, 291, 377, 379, 380, 382, 384, 385, 386, 388, 389, 390, 391, 398, 399, 400, 401, 402, 408, 409, 410, 411, 413, 414, 415, 416, 417, 419, 420, 428, 429, 430, 431

See also Chen et al. (2007), 89; Liu et al. (2013), 22, 382, 390, 391, 399, 400, 401, 428, 429, 430; Chuenwattanapranithi et al. (2008), 388, 401; Liu et al. (2013), 8, 382, 387, 390, 391, 399, 400, 428, 429, 430; Prom-On et al. (2016), 34

Yates, A. *See* Glewwe et al. (2018), 15
Yip, M. J., 7
Ylitalo, Rikka. *See* Nakai et al. (2009), 38, 242; Nakai et al. (2012), 242
Yoo, 164–165
Yoon, T., 368, 369
  See also Hasegawa-Johnson et al. (2005), 369
Yoruba language, final lowering in, 42
Yu, A. C. L, 8
  See also Liu et al. (2013), 22; Liu et al. (2013), 8, 382, 387, 390, 391, 399, 400, 428, 429, 430
Yuan, J., 388
Yuen, Ivan, 42

Zarco, W., 242
  See also Merchant et al. (2008), 242
Zec, D., 192
Zelaznik, H. N., 241, 242
  See also Schmidt et al. (1979), 242
Zellers, M., 297, 298
Zemlin, W. R., 384
Zeng, A. *See* Glewwe et al. (2018), 15
Zhang, C. *See* Peng et al. (2012), 183
Zhang, J., 388
Zhang, T. *See* Hasegawa-Johnson et al. (2005), 369
Zheng, H.-Y. *See* Peng et al. (2012), 183
Zhi, N., 139
Zue, Victor, 18
Zygis, Marzena, 26, 35, 39, 44–45
Zymet, J. *See* Glewwe et al. (2018), 15